Statistical Learning for Data Science

Lecture 09

唐晓颖

电子与电气工程系

南方科技大学

March 20, 2023

Multiple Linear Regression

1. Is there a relationship between the response and predictors?

- *The approach of using an F-statistic to test for any association between the predictors and the response works when p is relatively small, and certainly small compared to n.*

- *When p>n, we cannot fit the multiple linear regression model using least squares, so the F-statistic based approach does not work. This high-dimensional setting will be discussed in detail later.*

Multiple Linear Regression

2. Deciding on important variables

If we conclude on the basis of the F-statistic based p-value that at least one of the predictors is related to the response, then it is natural to wonder which are the ones!

After selecting the important variables, we can fit a single model involving only those predictors.

“Variable Selection”

Multiple Linear Regression

2. Deciding on important variables

Variable selection

Ideally, we would like to try out a lot of (and maybe all possibly) different models, each containing a different subset of the predictors, and then select the best one

- According to some criterion.

- Plotting various model outputs, such as the residuals, in order to search for patterns.

Multiple Linear Regression

2. Deciding on important variables

Variable selection

Ideally, we would like to try out a lot of (and maybe all possibly) different models, each containing a different subset of the predictors, and then select the best one.

Not practical! We need an **automated** and **efficient** approach to choose a smaller set of models to consider!

Multiple Linear Regression

2. Deciding on important variables

Variable selection

Forward selection

- Begin with the *null model* – a model that contains an intercept but no predictors.
- Fit *p* simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS among all new two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Multiple Linear Regression

2. Deciding on important variables

Variable selection

Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value – that is, the variable that is the least statistically significant.
- The new (p-1)-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significant threshold.

Multiple Linear Regression

2. Deciding on important variables

Variable selection

Mixed selection (a combination of forward and backward selection)

- Start with no variables in the model, and then add the variable that provides the best fit (similar to forward selection).
- Add variables one-by-one until the p-value for one of the variables in the model rises above a certain threshold.
- Remove that variable from the model.
- Continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Multiple Linear Regression

2. Deciding on important variables

Variable selection

Comments on the three methods:

- Backward selection cannot be used if p>n.
- Forward selection can always be used.
- Forward selection might include variables early that later become redundant.

Multiple Linear Regression

2. Deciding on important variables

Variable selection

- Later we discuss more systematic criteria for choosing an “optimal” member in the path of models produced by forward and backward stepwise selection.
- These include Mallow’s C_p , Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2 statistic, and Cross-validation (CV).

Multiple Linear Regression

3. Assessing fit of the model

R^2 *Statistic (the fraction of variance explained)*

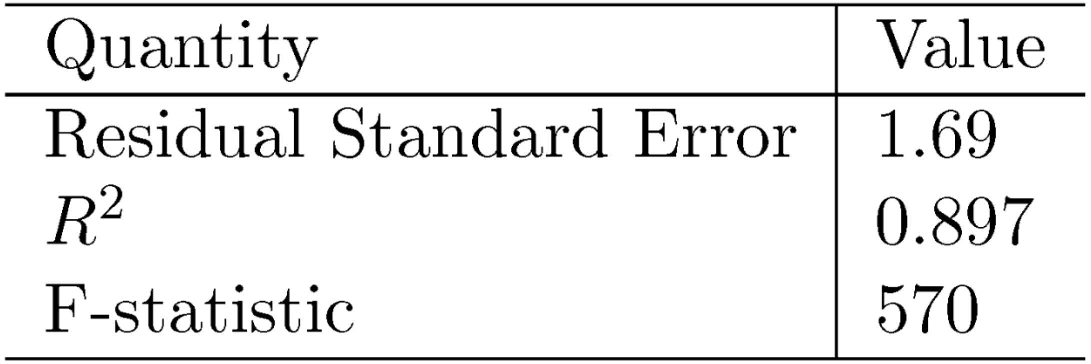
- In simple linear regression, we have $R^2 = [\mathrm{cor}(X, Y)]^2$
- In multiple linear regression, we have $R^2 = [\mathrm{cor}(Y, \hat{Y})]^2$, the square of the correlation between the response and the fitted linear model; in fact, one property of the fitted linear model is that it maximizes this correlation among all possible linear models.
- An R^2 value close to 1 indicates that the model explains a large portion of the variance in the response variable.

Multiple Linear Regression

3. Assessing fit of the model

R^2 *Statistic (the fraction of variance explained)*

Advertising data (sales versus TV, radio, and newspaper)



Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

$R^2 = 0.8972$

Multiple Linear Regression

3. Assessing fit of the model

R^2 *Statistic (the fraction of variance explained)*

Advertising data (sales versus TV and radio)

$$R^2 = 0.89719$$

- Even though the p-value for newspaper advertising is not significant (p=0.8599), there is still a small increase in R^2 if we include newspaper advertising in the model that already contains TV and radio advertising.

- The fact that there is only a tiny increase provides additional evidence that newspaper can be dropped from the model.

- Essentially, newspaper provides no real improvement in the model fit to the training samples, and its inclusion will likely lead to poor results on independent testing samples due to overfitting R^2要在变量个数相同的模型中去比较

Multiple Linear Regression

3. Assessing fit of the model

R^2 *Statistic (the fraction of variance explained)*

- R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.

- Adding another variable to the least squares equations allows us to fit the training data (though not necessarily the testing data) more accurately.

- The R^2 statistic, which is computed on the training data, must increase.

Multiple Linear Regression

3. Assessing fit of the model

R^2 *Statistic (the fraction of variance explained)*

Advertising data (sales versus TV and radio)

$$R^2 = 0.89719$$

Advertising data (sales versus TV)

$$R^2 = 0.61$$

- Adding radio to the model leads to a substantial improvement in R^2 .
- A model that uses TV and radio to predict sales is substantially better than the one that uses only TV.

Multiple Linear Regression

3. Assessing fit of the model

RSE

Advertising data (sales versus TV and radio)

$RSE = 1.681$

Advertising data (sales versus TV)

$RSE = 3.26$

This again shows that a model using TV and radio to predict sales is much more accurate (on the training data) than the one that only uses TV spending.

Multiple Linear Regression

3. Assessing fit of the model

RSE

Advertising data (sales versus TV and radio)

$RSE = 1.681$

Advertising data (sales versus TV, radio, and newspaper)

$RSE = 1.686$

This again shows that there is no point in also using newspaper spending as a predictor in the model.

Multiple Linear Regression

3. Assessing fit of the model

RSE

Advertising data (sales versus TV and radio)

$RSE = 1.681$

Advertising data (sales versus TV, radio, and newspaper)

$RSE = 1.686$

Why RSE increases when newspaper is added to the model given that RSS must decrease (more accurate)?

Multiple Linear Regression

3. Assessing fit of the model

RSE

$$RSE = \sqrt{\frac{1}{n-p-1} RSS}$$

Models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in p.

Multiple Linear Regression

4. Assessing accuracy of the prediction

- Confidence intervals for the coefficient estimates

- Model bias (the assumption of a linear model may be biased)

- Prediction intervals (a combination of both the error in the estimation (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error))

Multiple Linear Regression

4. Assessing accuracy of the prediction

The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, .., \hat{\beta}_p$ are estimates for $\beta_0, \beta_1, .., \beta_p$.

That is, the *least squares plane*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + .. + \hat{\beta}_p X_p$$

is only an estimate for the *true population regression plane*

$$f(X) = \beta_0 + \beta_1 X_1 + .. + \beta_p X_p$$

The inaccuracy in the coefficient estimates is related to the reducible error. We can compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$

Multiple Linear Regression

4. Assessing accuracy of the prediction

We use a confidence interval to quantify the uncertainty surrounding the average sales over a large number of cities.

For example, given that \$100,000 is spent on TV advertising and \$20,000 is spent on radio advertising in each city, the 95% confidence interval is [10,985, 11,528].

We interpret this to mean that 95% of intervals of this form will contain the true value of f(X).

In other words, if we collect a large number of data sets like the Advertising data set, and we construct a confidence interval for the average sales on the basis of each data set (given \$100,000 in TV and \$20,000 in radio advertising), then 95% of these confidence intervals will contain the true value of average sales.

Multiple Linear Regression

4. Assessing accuracy of the prediction

In practice assuming a linear model for f(X) is almost always an approximation of reality, so there is an additional source of potentially reducible error which we call *model bias*.

Multiple Linear Regression

4. Assessing accuracy of the prediction

Even if we knew f(X)—that is, even if we knew the true values for $\beta_0, \beta_1, .., \beta_p$ —the response value cannot be predicted perfectly because of the random error in the model (*irreducible error*).

How much will Y vary from $\hat{Y}$? We use *prediction intervals* to answer this question.

Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for f(X) (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

Multiple Linear Regression

4. Assessing accuracy of the prediction

A prediction interval can be used to quantify the prediction uncertainty surrounding sales for a particular city.

Given that $100,000 is spent on TV advertising and $20,000 is spent on radio advertising in that city the 95% prediction interval is [7,930, 14,580].

We interpret this to mean that 95% of intervals of this form will contain the true value of Y for this city.

Note that both intervals are centered at 11,256, but that the prediction interval is substantially wider than the confidence interval, reflecting the increased uncertainty about sales for a given city in comparison to the average sales over many locations.

Other considerations in the regression model

- Considerations in linear regression
  - Linear relationship
  - Variable selection
  - Data quality
  - Model evaluation
  - Model interpretation

Other considerations in the regression model

- Considerations in linear regression
  - Linear relationship
  - Variable selection
  - Data quality
  - Model evaluation
  - Model interpretation

Qualitative predictors

Model extension

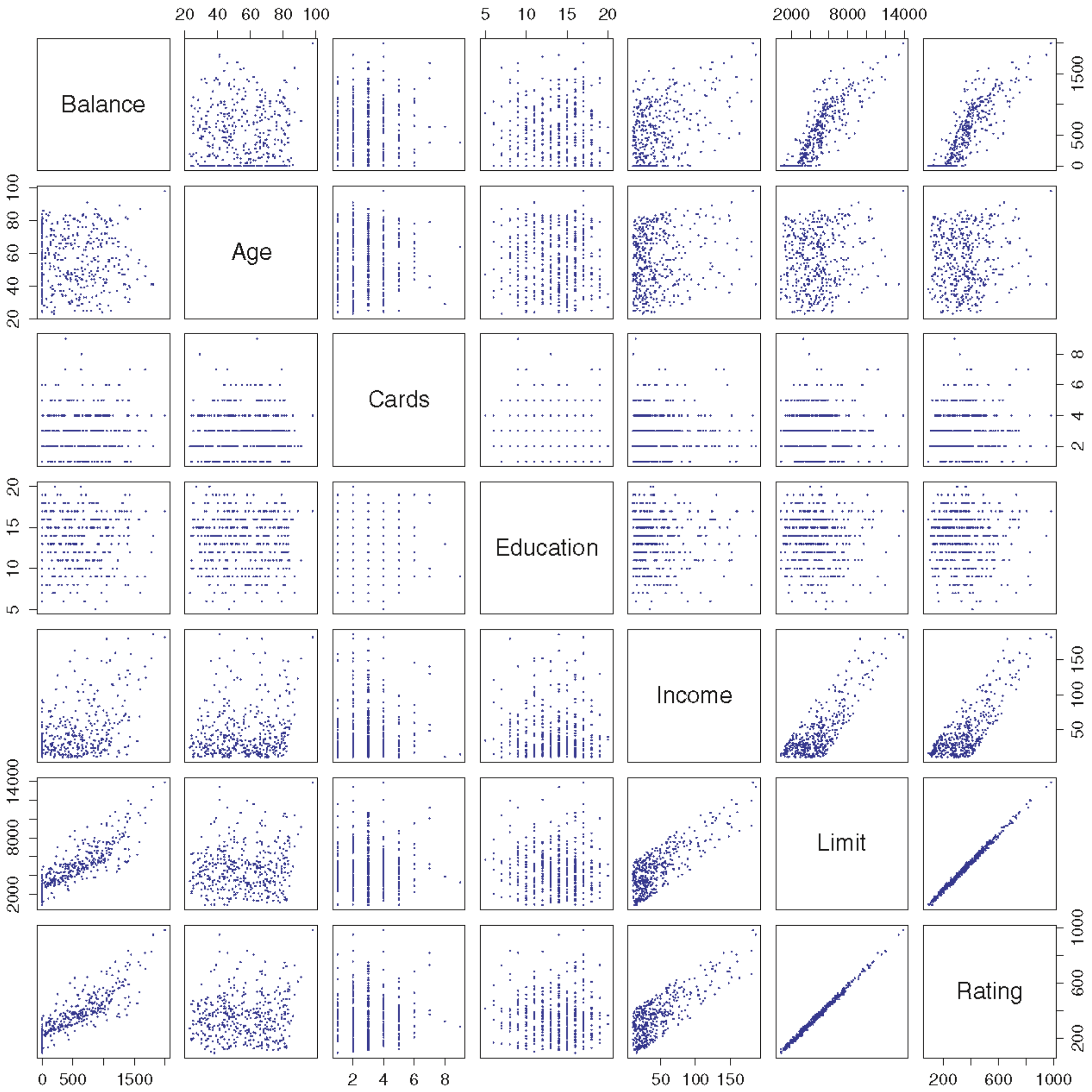
Other considerations in the regression model

- Qualitative predictors
  - Some predictors are not quantitative but are qualitative, taking a discrete set of values.
  - These are also called *categorical predictors* or *factor variables*.
  - Example: geographical position of a house in a house price prediction model
  - In regression problems, qualitative predictors usually need to be converted into numerical variables to be used by the model.

Other considerations in the regression model

- Qualitative predictors

In addition to the 7 quantitative variables, we also have four qualitative variables: gender, student (student status), marital status (married or single), and ethnicity (Caucasian, African American, or Asian).



Other considerations in the regression model

- Qualitative predictors

Predictors with only two levels

Investigate differences in credit card balance between males and females, ignoring the other variables

- Create an indicator or dummy variable that takes on two possible numerical values.

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Other considerations in the regression model

- Qualitative predictors

Predictors with only two levels

Investigate differences in credit card balance between males and females, ignoring the other variables

- Use this dummy variable as a predictor in the regression equation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Other considerations in the regression model

- Qualitative predictors

Predictors with only two levels

Investigate differences in credit card balance between males and females, ignoring the other variables

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

- β_0 -- the average credit card balance among males
- $\beta_0 + \beta_1$ -- the average credit card balance among females
- β_1 -- the average difference in credit card balance between females and males

Other considerations in the regression model

- Qualitative predictors

Predictors with only two levels

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	<0.0001
Gender [Female]	19.73	46.05	0.429	0.6690

The p-value for the dummy variable is very high. This indicates that there is no statistical evidence of a difference in average credit card balance between the genders.

Other considerations in the regression model

- Qualitative predictors

Predictors with only two levels

- The decision to code females as 1 and males as 0 is arbitrary, and has no effect on the regression fit.

- Instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

Other considerations in the regression model

- Qualitative predictors

Predictors with only two levels

- Instead of a 0/1 coding scheme, we could create a dummy variable

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

- β_0 -- the overall average credit card balance (ignoring the gender effects)

- β_1 -- the amount that females are above the average and males are below the average

Other considerations in the regression model

- Qualitative predictors

Predictors with only two levels

Note: the final predictions for the credit balances of males and females will be identical regardless of the coding scheme used. The only difference is in the way that the coefficients are interpreted.

Other considerations in the regression model

- Qualitative predictors

Predictors with more than two levels

Investigate differences in credit card balance among different ethnicity groups.

- Create additional dummy variables

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

Other considerations in the regression model

- Qualitative predictors

Predictors with more than two levels

Investigate differences in credit card balance among different ethnicity groups.

- Use these variables in the regression equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

β_0 -- the average credit card balance for African Americans

β_1 -- the difference in the average balance between Asian and African American categories

β_2 -- the difference in the average balance between Caucasian and African American categories

Other considerations in the regression model

- Qualitative predictors

Predictors with more than two levels

Note: there will always be one fewer dummy variables than the number of levels.

Other considerations in the regression model

- Qualitative predictors

Predictors with more than two levels

Investigate differences in credit card balance among different ethnicity groups.

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	<0.0001
Ethnicity [Asian]	-18.96	65.02	-0.287	0.7740
Ethnicity [Caucasian]	-12.50	56.68	-0.221	0.826

Other considerations in the regression model

- Qualitative predictors

Predictors with more than two levels

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	<0.0001
Ethnicity [Asian]	-18.96	65.02	-0.287	0.7740
Ethnicity [Caucasian]	-12.50	56.68	-0.221	0.826

Investigate differences in credit card balance among different ethnicity groups.

- The estimated balance for African American is $531.00.
- The Asian category will have $18.69 less debt than the African American category.
- The Caucasian category will have $12.50 less debt than the African American category.
- The p-values associated with the coefficient estimates for the two dummy variables are very large, suggesting no statistical evidence of a real difference in credit card balance between the ethnicities.

Other considerations in the regression model

- Qualitative predictors

Predictors with more than two levels

Investigate differences in credit card balance among different ethnicity groups.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

Alternatively, we can use an F-test to perform hypothesis testing $H_0 : \beta_1 = \beta_2 = 0$

- This hypothesis testing does not depend on the coding
- This F-test has a p-value of 0.96, indicating that we cannot reject the null hypothesis that there is no relationship between balance and ethnicity

Other considerations in the regression model

- Qualitative predictors

  Note:

  - This dummy variable approach also works when there are both quantitative and qualitative predictors.

  - There are many different ways of coding qualitative variables besides the dummy variable approach taken here. They all lead to equivalent model fits, but the coefficients are different and have different interpretations.

Other considerations in the regression model

- Extensions of the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- Why do we need extension?
  - In linear regression, we assume that the relationship between the predictor variables and the response variable is linear.
  - However, in many real-life situations, a linear model may not adequately capture the underlying relationship between the variables.
- The extension is to improve the predictive performance of the linear model.

Other considerations in the regression model

- Extensions of the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- Extensions of linear model
  - Examples: Polynomial regression, generalized linear models, kernel regression, …
  - Help better fit the data to improve the predictive ability of the model
  - Consider about overfitting and underfitting
- Removing additive assumptions of linear model
  - Allowing for non-additive effects of predictor variables on the response variable, such as interactions between predictors or nonlinear relationships.
  - Does not necessarily mean that the effect of each predictor variable is dependent on the values of the other predictor variables.

Other considerations in the regression model

- Extensions of the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Two highly restrictive assumptions:

- Additive: the effect of changes in the predictor on the response is independent of the values of the other predictors (there may be interactions).

- Linear: the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j (there may be nonlinearity).

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

- In our previous analysis of the advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.

- For example, the linear model

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \varepsilon$$

states that the average effect on sales of a one-unit increase in TV is always β_1 regardless of the amount spent on radio.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

- But suppose that spending money on radio advertising actually increase the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.

- In this situation, given a fixed budget of $100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.

- In marketing, this is known as a ***synergy*** effect, and in statistics it is referred to as an ***interaction*** effect.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

The standard linear regression model with two variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Note: if we increase X_1 by one unit, then Y will increase by an average of β_1 units. The presence of X_2 does not alter this statement.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Extending the model to allow for interaction effects:

The standard linear regression model with two variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underbrace{\beta_3 X_1 X_2}_{\textit{Interaction term}} + \varepsilon$$

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Extending the model to allow for interaction effects:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \\ &= \beta_0 + \underbrace{(\beta_1 + \beta_3 X_2)}_{} X_1 + \beta_2 X_2 + \varepsilon \end{aligned}$$

The effect of X_1 on Y is no longer constant:
adjusting X_2 will change the impact of X_1 on Y .

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: To predict the number of units produced on the basis of the number of production lines and the total number of workers.

- The effect of increasing the number of production lines will depend on the number of workers, since it no workers are available to operate the lines, then increasing the number of lines will not increase production.

- It would be appropriate to include an interaction term between lines and workers in a linear model to predict units.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: To predict the number of units produced on the basis of the number of production lines and the total number of workers.

Suppose that when we fit the model, we obtain

$$\begin{aligned}\text{units} &\approx 1.2 + 3.4 \times \text{lines} + 0.22 \times \text{workers} + 1.4 \times (\text{lines} \times \text{workers}) \\ &= 1.2 + (3.4 + 1.4 \times \text{workers}) \times \text{lines} + 0.22 \times \text{workers}\end{aligned}$$

Conclusion: adding an additional line will increase the number of units produced by $3.4 + 1.4 \times \text{workers}$. Hence, the more workers we have, the stronger will be the effect of lines.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

$$
\begin{aligned}
\text{sales} &\approx \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{TV} \times \text{radio}) + \varepsilon \\
&= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \varepsilon
\end{aligned}
$$

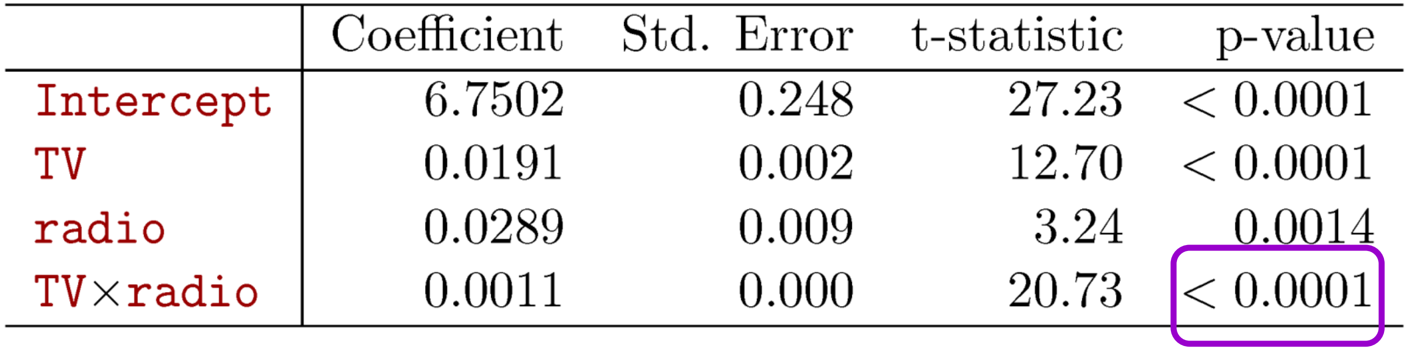
β_3 : the increase in the effectiveness of TV advertising for a one unit increase in radio advertising (or vice-versa).

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)



	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Conclusion: the p-value for the interaction term, $\mathrm{TV} \times \mathrm{radio}$, is extremely low, indicating that there is strong evidence for $H_a : \beta_3 \neq 0$. In other words, it is clear that the true relationship is not additive.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

R^2 *Statistic (the fraction of variance explained)*

With interaction term (sales versus TV and radio)

$$R^2 = 0.968$$

Without interaction term (sales versus TV and radio)

$$R^2 = 0.897$$

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

R^2 *Statistic (0.968 versus 0.897)*

Conclusion:

- The model that includes the interaction term is superior to the model that contains only main effects.

- (96.8-89.7)/(100-89.7) = 69% of the variability in sales that remains after fitting the additive model has been explained by the interaction term.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

$$\text{sales} = 6.7502 + 0.0191 \times \text{TV} + 0.0289 \times \text{radio} + 0.0011 \times (\text{TV} \times \text{radio}) + \varepsilon$$

Conclusion:

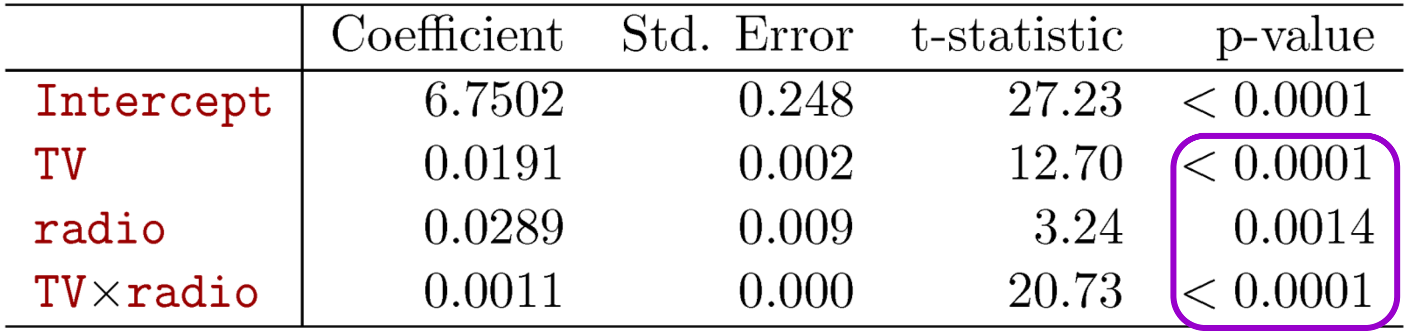
- An increase in TV advertising of $1,000 is associated with increased sales of $(\beta_1 + \beta_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio}$ units.
- An increase in radio advertising of $1,000 is associated with increased sales of $(\beta_2 + \beta_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV}$ units.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)



	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Conclusion: both the main effects (TV, radio) and the interaction effect (TV×radio) are statistically significant, so it is obvious that all three variables should be included in the model.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Question: What if an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not?

Answer: The hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant. **(Why?)**

Rationale: If $X_1 \times X_2$ is related to the response, then whether or not the coefficients of X_1 or X_2 are exactly zeros is of little interest. Also $X_1 \times X_2$ is typically correlated with X_1 and X_2, and so leaving them out tends to alter the meaning of the interaction.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Interactions of qualitative variables, or a combination of quantitative and qualitative variables.

Example: Credit data (balance versus income (quantitative) & student (qualitative))

No interaction term:

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2, & \text{if } i\text{th person is a student} \\ 0, & \text{if } i\text{th person is not a student} \end{cases}$$

$$= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2, & \text{if } i\text{th person is a student} \\ \beta_0, & \text{if } i\text{th person is not a student} \end{cases}$$

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Credit data (balance versus income (quantitative) & student (qualitative))

No interaction term:

$$\text{balance}_i = \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2, & \text{if } i\text{th person is a student} \\ \beta_0, & \text{if } i\text{th person is not a student} \end{cases}$$

- Fitting two lines, one for students and one for non-students
- The two lines have different intercepts, $\beta_0 + \beta_2$ versus β_0 , but the same slope,β_1