Predictive Modeling

Multiple Linear Regression, Qualitative Predictors and Interaction Effects

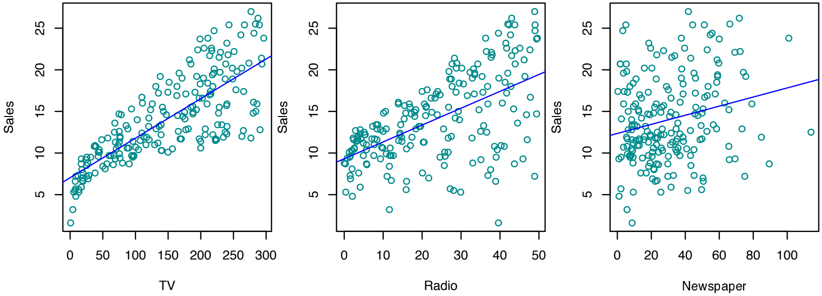
Mirko Birbaumer

HSLU T&A

Multiple Linear Regression: Example `Advertising`

`Advertising` example: How can we predict `sales` on the basis of advertising expenditures in `TV`, `radio` and `newspaper`?

3 separate simple regression models:



Multiple Linear Regression: Example Advertising

	Coefficient	Std.error	t-statistic	p-value
Intercept	7.033	0.458	15.36	< 0.0001
TV	0.048	0.003	17.67	< 0.0001

	Coefficient	Std.error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
Radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std.error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
Newspaper	0.055	0.017	3.30	< 0.0001

Tabelle: Simple linear regression models for each advertising medium

Extending the Simple Linear Regression Model

Critical questions concerning the separate simple linear regression models:

- How to make a prediction of `sales` given levels of the three advertising media budgets?

- Each of the three regression equations ignores the other two media in forming estimates for the regression coefficients

Multiple Linear Regression

Multiple Linear Regression Model

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

- X_j : j^{th} predictor variable
- β_j : association between X_j and response variable Y

Examples: Multiple Linear Regression

Example 1: `Advertising`

$$\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper} + \varepsilon$$

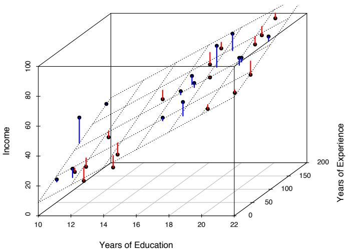
$\longrightarrow$ graphical representation impossible

Example 2: `Income`

$$\texttt{income} = \beta_0 + \beta_1 \cdot \texttt{education} + \beta_2 \cdot \texttt{experience} + \varepsilon$$

$\longrightarrow$ graphical representation is possible, since 2 predictors and response variable can be visualized

Example 2: Income



Parameter Estimation: Minimize RSS

$$\mathrm{RSS} = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip})^2$$

Result: See Example 1.3 in the Multiple Linear Regression chapter

$$\texttt{Income} \approx -50.086 + 5.896 \cdot \texttt{education} + 0.173 \cdot \texttt{experience}$$

Example: Advertising

See Example 2.1 in the Multiple Linear Regression chapter how the following multiple regression coefficients are determined

	Coefficient	Std.error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
Radio	0.189	0.0086	21.89	< 0.0001
Newspaper	-0.001	0.0059	-0.18	0.8599

- Comparing these coefficients to those obtained through simple linear regression, we observe: multiple linear regression coefficient estimates for TV and Radio are similar to the simple linear regression coefficient estimates
- However, $\hat{\beta}_3$ for newspaper is different from 0 in the simple linear regression model, whereas in multiple linear regression it is approximately 0

Example: Advertising

How to solve for this contradiction: Correlation matrix

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

- Correlation between radio and newspaper: 0.35
- The more money is spent on radio, the higher is the advertising budget for newspaper
- Higher advertising budgets for newspaper tend to be associated with higher values of radio due to their positive correlation, but radio is actually the predictor that **influences** sales

Some Important Questions in Multiple Linear Regression

1. *Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?*

2. *Do all the predictors $X_1, \ldots, X_p$ help to explain Y, or is only a subset of the predictors useful?*

3. *How well does the model fit the data?*

4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

1. Is there a Relationship Between the Response and Predictors?

- **Simple Linear Regression**: If $\beta_1 = 0$, then there is no relationship between predictor and response variable, otherwise we would conclude, that there is a relationship

- **Multiple Linear Regression**: Are **all** regression coefficients - with exception of β_0 - zero? We test the null hypothesis:

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

versus the alternative hypothesis

$$H_A: \quad \text{at least one } \beta_i \text{ is non-zero}$$

Hypothesis Test using the F-statistic

- The hypothesis test is performed by computing the **F-statistic**

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ and $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- If linear model assumptions are **correct**, it can be shown

$$\mathrm{E}\left[\frac{\text{RSS}}{n-p-1}\right] = \sigma^2$$

- Provided the null hypothesis is **true**, then it can be shown

$$\mathrm{E}\left[\frac{\text{TSS - RSS}}{p}\right] = \sigma^2$$

- If there is **no** relationship between the response and the predictors: value of F-statistic approximately **1**

Hypothesis Test using the F-statistic

- If H_A is **true**, then
$$\mathrm{E}\left[\frac{\mathsf{TSS} - \mathsf{RSS}}{p}\right] > \sigma^2$$
and we expect F to be greater than 1

- When H_0 is **true** and the errors ε_i follow a normal distribution, the F-statistic follows an F-**distribution** with p and $n - p - 1$ degrees of freedom

- For any given value of n and p, using the F-distribution the p-value associated with the F-statistic can be computed

Hypothesis Test using the F-statistic: `Advertising` Example

- **Example**: Multiple Linear Model for `Advertising` data set
- Value of `F-statistic` is 570, see example `3.1` in the `Multiple Linear Regression` chapter
- `p-value` associated with this F-statistic is essentially **zero**
- **Conclusion**: at least **one** of the media is associated with `sales`

2. Deciding on Important Variables

- The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as **variable selection**
- Variable selection is studied in Chapter *Linear Model Selection*

3. How well does the model fit the data?

Two of the most common numerical measures of **model fit**:

- RSE : **Residual Standard Error**
- R^2 : **fraction of variance** explained by the regression model
  - In multiple linear regression: $R^2 = \mathrm{Cor}[Y, \hat{Y}]^2 \;\rightarrow\;$ the square of the correlation between response and predicted response
  - An R^2 value close to 1 indicates that the model explains a **large** fraction of the variance in the response variable

3. How well does the model fit the data?

Example: `Advertising`

- Linear model with predictors `TV` , `newspaper` and `radio` for the response `sales`: $R^2 = 0.8972$

- See example `3.1` in the `Multiple Linear Regression` chapter

- Linear model with predictors `TV` and `radio` for the response `sales`: $R^2 = 0.89719$

Conclusion: **very small** increase in R^2 if we include `newspaper` in the multiple linear regression model

4. Prediction: Confidence Interval for the true average `sales`

- Given that CHF 100 000 is spent on `TV` advertising and CHF 20 000 is spent on `radio` advertising in each city, the 95 % confidence interval for the true average `sales`:

$$[10'985, 11'528]$$

- See Example 3.5 in the `Multiple Linear Regression` chapter

4. Prediction: Prediction Interval for `sales`

- Given that CHF 100 000 is spent on `TV` advertising and CHF 20 000 for `radio` advertising in a **particular** city, the 95 % **prediction interval** for `sales` in that city

$$[7'930, 14'583]$$

- See Example 3.5 in the Multiple Linear Regression chapter

Variety of Regression Modeling

Example: `Advertising`

Large model:

$$\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper} + \varepsilon$$

$R^2 = 0.8972$ and p-value for $H_0 : \beta_3 = 0$ is 0.8599

Small modell:

$$\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \varepsilon$$

$R^2 = 0.89719$

Questions:

- How can we compare the *large* model with the *small* model?
- How can we test that a particular subset of q of the coefficients are zero?

Repetition: F-Test

F-Test:

- Null hypothesis

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

- Alternative hypothesis

$$H_A: \quad \text{at least one } \beta_i \text{ is non-zero}$$

- Distribution of F-statistic assuming H_0 is true

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim \mathcal{F}_{p,n-p-1}$$

- If H_0 is true, then F-statistic has a value near to 1
- If H_A is true, then the value of the F-statistic is larger than 1

F-Test for a Subset Consisting of q Predictors

- **Question**: How can we test whether a subset consisting of q coefficients is zero?

- **Null hypothesis** (subset of q coefficients β_i are zero):

$$H_0 : \quad \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$$

F-Test for a Subset Consisting of q Predictors

- **F-statistic**:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$$

  RSS_0 : residual sum of squares for the *small* model, for which q predictor variables were omitted

- **Distribution** of the test statistic F assuming H_0 is **true**

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

- If H_0 is **true**, then the value of the F-statistic is approximately 1, otherwise it is larger than 1

- If p-value associated with F-statistic is smaller than the significance level α, then we **reject** the null hypothesis

Anova - Analysis of Variance

- See Example `4.1` in the `Multiple Linear Regression` chapter
- Residual sum of squares (`RSS`) in the *large* model, resp. in the *small* model

$$\mathrm{RSS} = 556.83 \qquad \mathrm{RSS}_0 = 556.91$$

- Degrees of freedom (`Res.DF`) in the *large* , resp. in the *small* model

$$n - p - 1 = 200 - 3 - 1 = 196 \quad ; \quad n - p_0 - 1 = 200 - 2 - 1 = 197$$

Anova - Analysis of Variance

- **F-statistic**:

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)} = \frac{(556.91 - 556.83)/1}{556.83/(200-3-1)} = 0.0312$$

- **p-value** `Pr(>F)` for this value of the F-statistic assuming the null hypothesis $\beta_3 = 0$ is **true** yields: 0.8599

- We retain the null hypothesis: predictor variable `newspaper` is redundant

- **Note**: F-test with $q = 1$ corresponds to t-test, when all other variables are considered in the model!

Qualitative Predictor Variables - Example `Credit`

Data set `Credit` was recorded in the USA:

- **Response Variable** : `balance` : average credit card debt for a number of individuals
- **Quantitative predictor variables**:
  - `age`
  - `cards` : number of credit cards
  - `education` : years of education
  - `income` : income in thousand of dollars
  - `limit` : credit card limit
  - `rating` : credit rating
- **Qualitative predictor variables (factors)**:
  - `gender`
  - `student` : student status
  - `ethnicity` : Caucasian, African American or Asian

Question: How can we incorporate **qualitative predictor variables** into a regression model?

Factor Variables with Two Levels - Example `Credits`

Goal: We wish to investigate differences in credit card `balance` between males und females (`gender`)

Solution: based on the `gender` variable, we create a new variable:

- **Indicator** or **dummy variable**:

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

- Regression model:

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

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

Factor Variables with Two Levels - Example `Credits`

- See examples 4.4 - 4.7 in the **Multiple Linear Regression** chapter

- **Interpretation**:
  - estimated average credit card debt for males: \$ 509.80
  - estimated average difference to females: \$ 19.73
  - estimated average debt for females: \$ 509.80 + \$ 19.73 = \$ 529.53
  - p-value for the dummy variable β_1 is 0.6690 : no statistical evidence of a difference in average credit card balance between genders

Factor Variables with Three Levels - Example `Credits`

The `ethnicity` variable has **three** possible levels. We require **two** dummy variables

- First dummy variable:

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

- Second dummy variable:

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

Regression model:

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

Factor Variables with Three Levels - Example `Credits`

- See example 4.8 in the `Multiple Linear Regression` chapter
- $\beta_0 = 531.00$: average credit card balance for **African Americans**
- $\beta_1 = -18.69$: difference in the average balance between the **Asian** and **African American** categories
- $\beta_2 = -12.50$: difference in the average balance between the **Caucasian** and **African American**

Regression Models with Factor Variables

- The level with no dummy variable - African American in the example - is known as the **baseline**

- Asian category will have \$ 18.69 less debt than African Americans category; Caucasians category will have \$ 12.50 less debt than African Americans

- However, the p-values associated with the coefficient estimates for the two dummy variables are very **large**: **no** statistical evidence of a real difference in credit card balance between ethnicities

Regression Models with Factor Variables

General Remarks:

- There is always one fewer dummy variable than the number of levels

- There are many different ways of **coding qualitative variables**, no effect on regression fit, but does alter interpretation of coefficients and p-values

Extensions of the Linear Model: Additivity versus Interaction

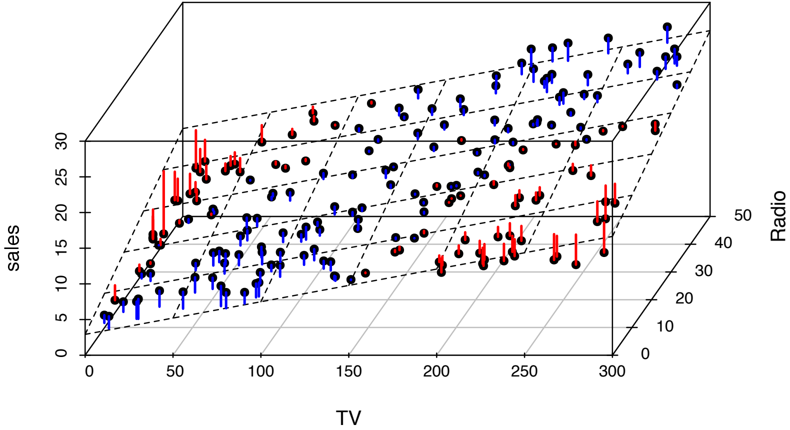
Additivity:
The effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors

Example:

$$\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \varepsilon$$

Additivity: average effect on `sales` of a one-unit increase in `TV` is always β_1 regardless of the amount spent on `radio`

Objection: suppose that spending money on `radio` advertising actually increases the effectiveness of `TV` advertising, so that the slope term for `TV` should increase as `radio` increases $\Rightarrow$ **Interaction Effect**



Spending half on `radio` and half on `TV` increases `sales` more than allocating the entire amount to either `TV` or to `radio` $\Rightarrow$ **Interaction effect**

Example `Advertising` with Interaction Term

A linear model that uses `TV`, `radio` and an **interaction** between the two to predict `sales` takes the form

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

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

- **Interaction term**: see example `4.11` of `Multiple Linear Regression` chapter

- **Interaction term**: see example 4.11 of `Multiple Linear Regression` chapter

- R^2 **with** interaction term: 0.968 ; R^2 **without** interaction term: 0.897

- Increase in `TV` advertising of CHF 1000 is associated with an increase in `sales` of (in units)

$$(\hat{\beta}_1 + \hat{\beta}_3 \cdot \texttt{radio}) \cdot 1.000 = 19 + 1.1 \cdot \texttt{radio}$$

- p-values associated with `TV`, `radio` and $\texttt{TV} \cdot \texttt{Radio}$ are all statistically significant

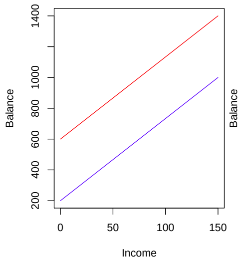
Example Credit without Interaction Term

We wish to predict balance using the income (quantitative) and student (qualitative) variables:

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

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

Model describes two parallel regression lines, one for students (red) and one for non-students (blue). Slope β_1 is the same, but the intercepts are different: $\beta_0 + \beta_2$ and β_0



Example `Credit` with Interaction Term

$$
\texttt{balance}_i \approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \cdot \texttt{income}_i & \text{if student} \\ 0 & \text{if not a student} \end{cases}
$$

$$
= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \cdot \texttt{income}_i & \text{if not a student} \end{cases}
$$

Two regression lines for students (red) and for non-students (blue) with **different slopes** $\beta_1 + \beta_3$ and β_1 in addition to different intercepts $\beta_0 + \beta_2$ and β_0