SDU Summer School

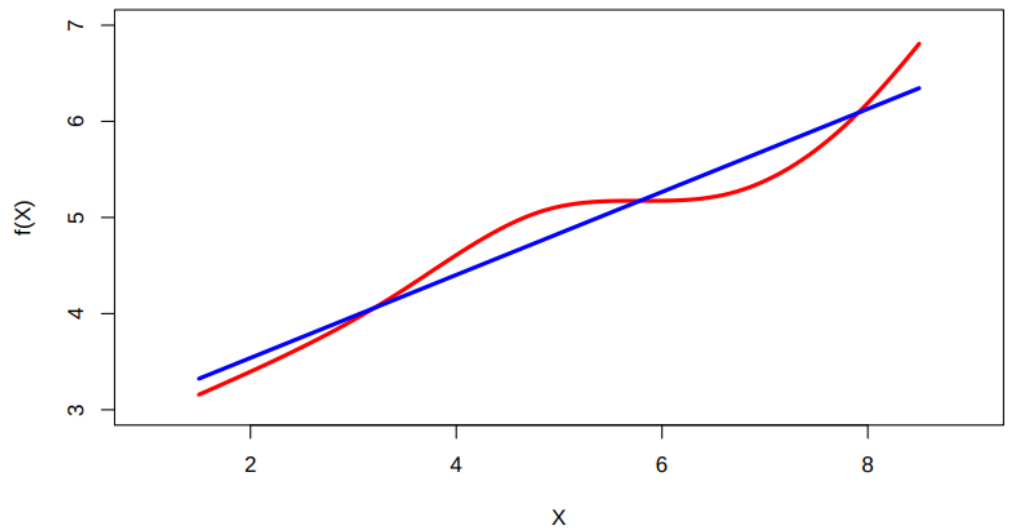
Deep Learning

Summer 2018

Linear Regression

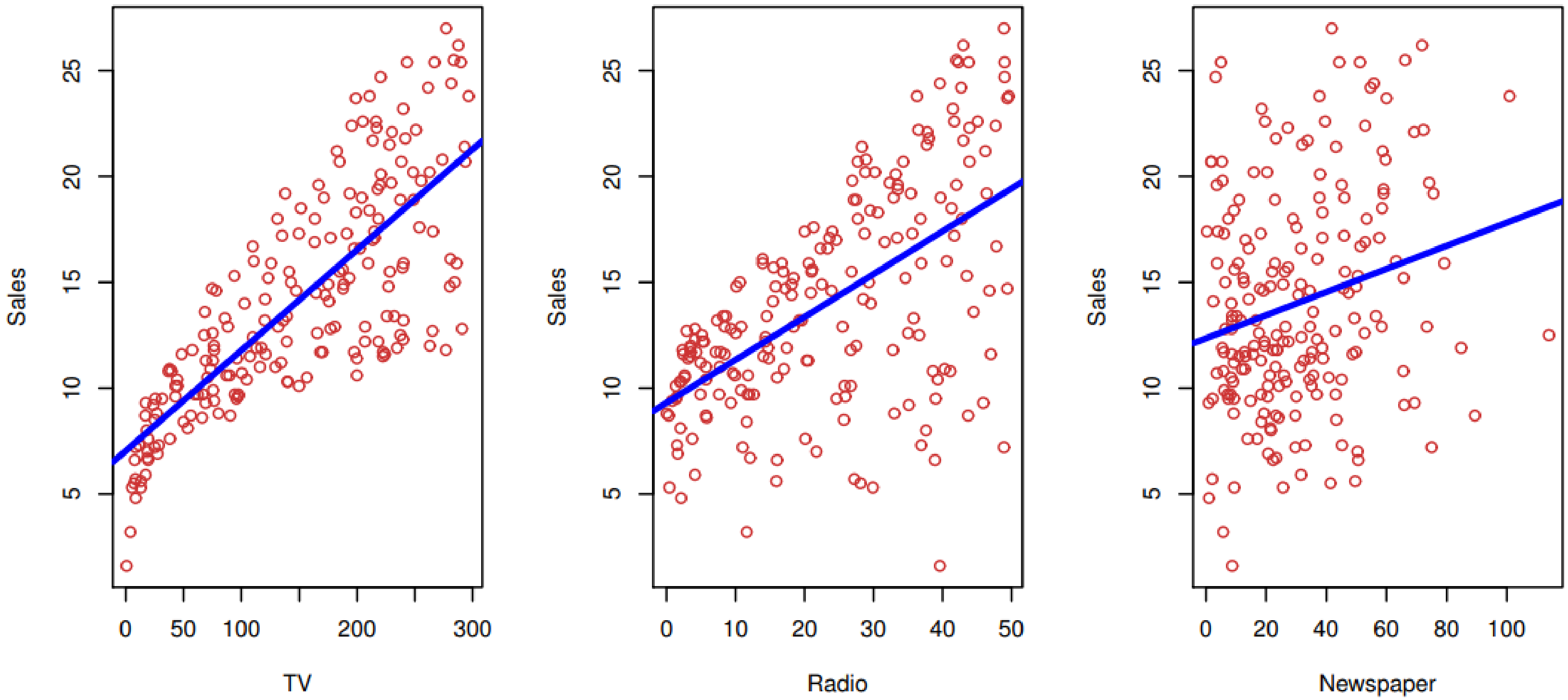
Linear Regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots X_p$ is linear.
- True regression functions are never linear!



- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Example: Advertising data



- The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

Example: Advertising data

Questions we might ask:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Simple Linear Regression.

- We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- where β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients or parameters, and ϵ is the error term.

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$model coefficients, we **predict** future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Estimation of the Parameters by Least Squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction of Y for the ith observation.
- Then the **residual** is defined as

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

- We define the **residual sum of squares** as

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2 =$$
$$\left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$

- Our task is to find the $\hat{\beta}_0$ and $\hat{\beta}_1$ minimizing the RSS.

Estimation of the Parameters by Least Squares

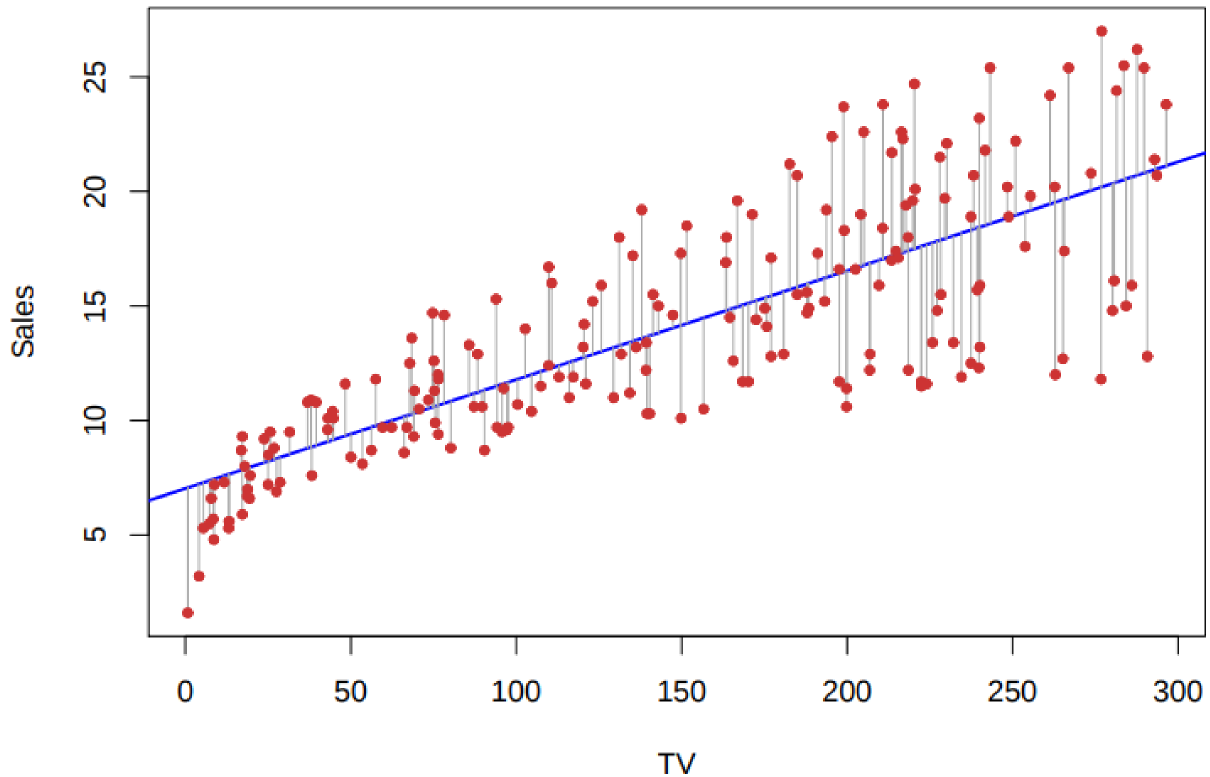
- With simple derivation of the RSS formula, we can show that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- with the sample means $\bar{y} = \frac{1}{2}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{2}\sum_{i=1}^{n} x_i$

Example: Advertising data



- The least squares fit for the regression of sales onto TV.
- In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Assessing the Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under repeated sampling.

- For our example, we have:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Confidence Intervals

- These standard errors can be used to compute confidence intervals.
- A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- It has the form

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

- In other words: there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

will contain the true value of $\hat{\beta}_1$.

Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients.

- The most common hypothesis test involves testing the **null hypothesis** of:

$\boldsymbol{H_0}$: There is no relationship between X and Y

$$H_0: \beta_1 = 0$$

versus the alternative hypothesis

$\boldsymbol{H_A}$: There is some relationship between X and Y

$$H_0: \beta_1 \neq 0$$

Hypothesis testing

- To test the null hypothesis, we compute a **t-statistic**, given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.

- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the **p-value**.

Assessing the Overall Accuracy

- We can compute the **Residual Sum of Squares** (RSS)

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

- With that, we can define the **Residual Standard Error** (RSE)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}}$$

Assessing the Overall Accuracy

- To assess how much of the variance is explained, we use the **R-squared** measure

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the **total sum of squares**.

- It can be shown that in this simple linear regression setting that $R^2 = r^2$ with r being the **correlation** between X and Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Multiple Linear Regression

- We can include several features or predictors into our model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + +\beta_p X_p + \epsilon$$

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

$$\text{sales} = \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{radio} + \beta_3 \cdot \text{newspaper} + \epsilon$$

In Matrix Form

- We can formulate the entire linear regression also in Matrixform

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

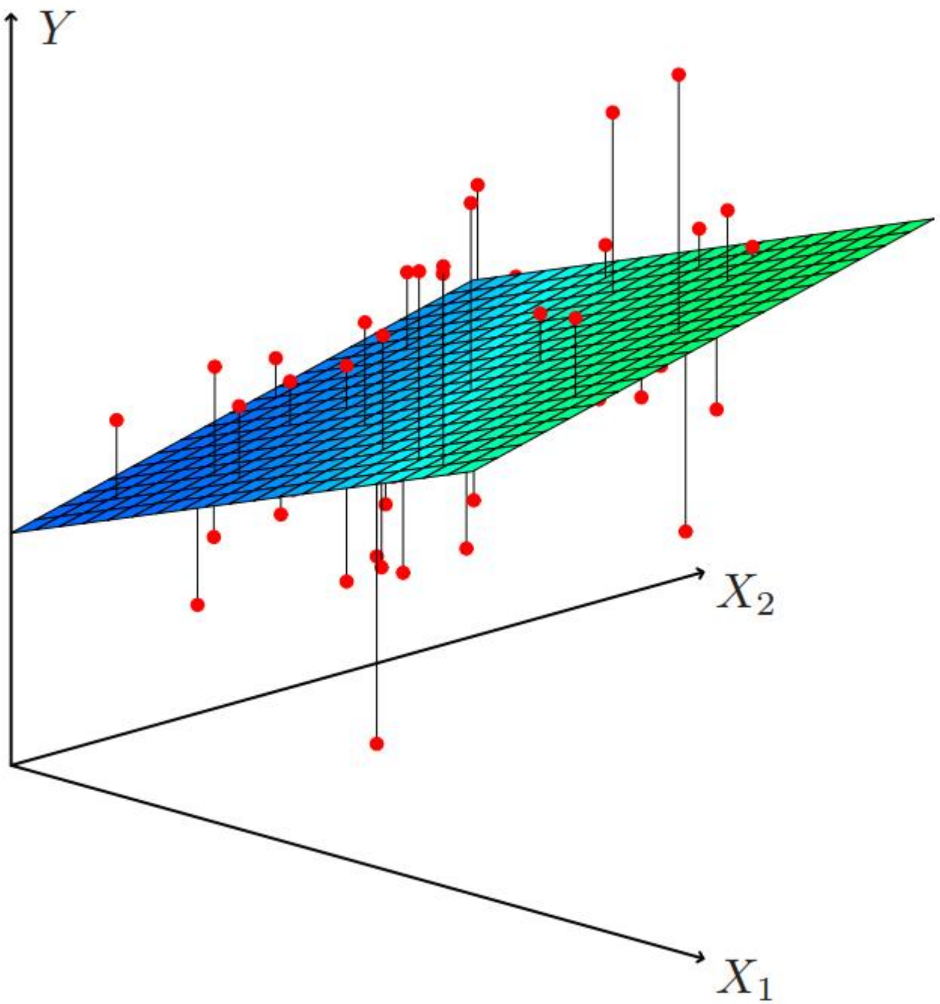
- $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ is the $n \times 1$ response vector
- $\boldsymbol{X} = [1_n, \boldsymbol{X}'] \in \mathbb{R}^{n\times(p+1)}$ is the $n \times (p+1)$ design matrix
  - 1_n is an $n \times 1$ vector of ones
  - $\boldsymbol{X}' = [\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p] \in \mathbb{R}^{n\times p}$ the $n \times p$ predictor Matrix
- $b = (\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}$ is regression coefficient vector
- $e = (e_1, e_2, \dots, e_3) \in \mathbb{R}^n$ is the $n \times 1$ error vector

Example

#	TV	Radio	News	Sales
1	230.1	37.8	69.2	**22.1**
2	44.5	39.3	45.1	**10.4**
3	17.2	45.9	69.3	**9.3**
4	151.5	41.3	58.5	**18.5**
5	180.8	10.8	58.4	**12.9**

$$\begin{pmatrix} 22.1 \\ 10.4 \\ 9.3 \\ 18.5 \\ 12.9 \end{pmatrix} = \begin{pmatrix} 1 & 230.1 & 37.8 & 69.2 \\ 1 & 44.5 & 39.3 & 45.1 \\ 1 & 17.2 & 45.9 & 69.3 \\ 1 & 151.5 & 41.3 & 58.5 \\ 1 & 180.8 & 10.8 & 58.4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix}$$

Linear Regression



Results for advertising data

	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

Correlations:

	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

Interpreting Regression Coefficients

"Data Analysis and Regression" Mosteller and Tukey 1977

- A regression coefficient β_j estimates the expected change in Y per unit change in X_j, with all other predictors held fixed. But predictors usually change together!

- Example: Y total amount of change in your pocket; X_1 = # of coins; X_2 = # of pennies, nickels and dimes. By itself, regression coefficient of Y on X_2 will be > 0. But how about with X_1 in model?

- Y = number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is $\hat{Y} = b_0 + 0.5 \cdot W - 0.1 \cdot H$. How do we interpret $\beta_2 < 0$?

Two quotes by famous Statisticians

"Essentially, all models are wrong, but some are useful"

George Box

"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"

Fred Mosteller and John Tukey, paraphrasing George Box

Some important questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?

2. Do all the predictors 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?

Is at least one predictor useful?

- For the first question, we can use the F-statistic

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

- It's similar to a T statistic from a T-Test; The T-test will tell you if a single variable is statistically significant and an F-test will tell you if a group of variables are jointly significant.

Deciding on the important variables

- The most direct approach is called **all subsets** or **best subsets regression**: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.

- However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models!

- Instead we need an automated approach that searches through a subset of them. We discuss two commonly use approaches next.

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 amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

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 significance threshold.

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**

Qualitative Predictors

- For Example: We have given a dataset with female and male. When we want to incorporate that into our model, we create a new dummy variable "is_female"

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

Resulting Model:

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

More than two levels

- If we have more categories than two, we simply create more dummy variables
- For example: eye_color = {blue, brown, green, grey}
- We create the dummy variables “has_blue_eyes”, “has_green_eyes”, “has_grey_eyes”.
- There is always one variable less than we have levels.
- The level with no dummy variable – brown eye color - is known as the baseline.

Extensions of the Linear Model

Removing the additive assumption: **interactions** and **nonlinearity**

Interactions:

- 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

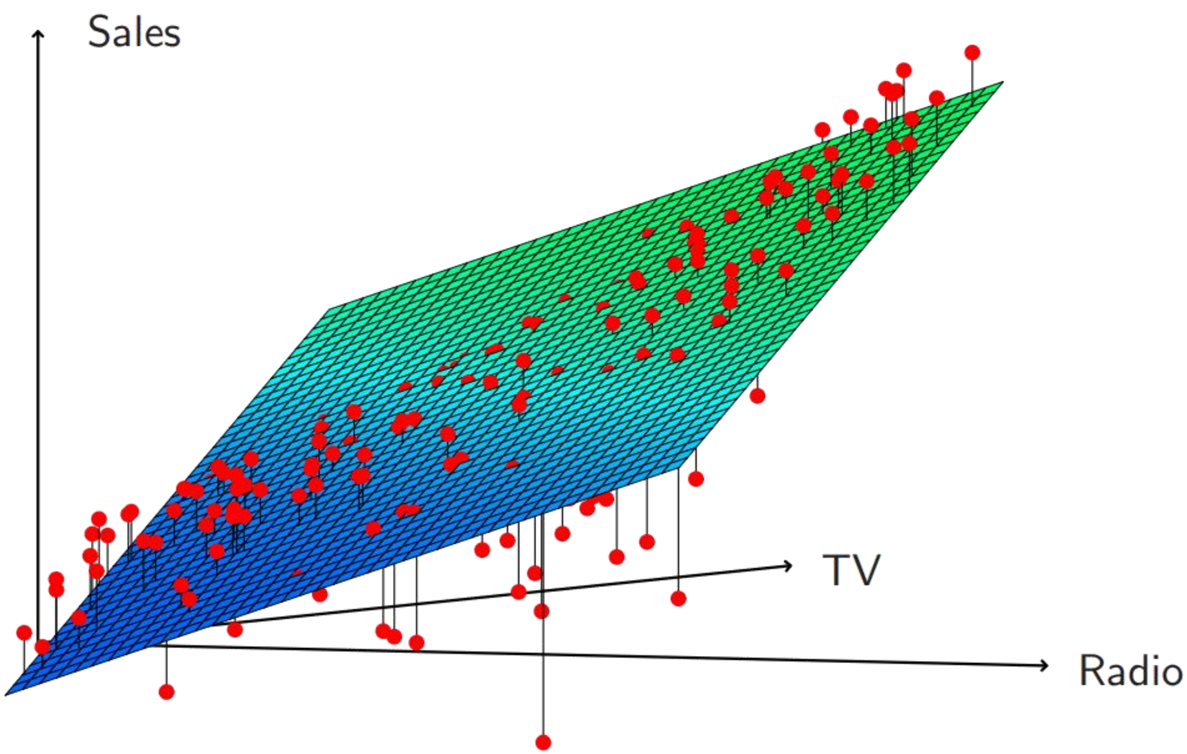
$$\text{sales} = \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{radio} + \beta_3 \cdot \text{newspaper} + \epsilon$$

  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.

Interactions

- But 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.
- 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.

Interaction in the Advertising data?



- When levels of either TV or radio are low, then the true sales are lower than predicted by the linear model.
- But when advertising is split between the two media, then the model tends to underestimate sales

New Model

Model takes the form

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

Results:

	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

Interpretation

- The results in this table suggests that interactions are important.
- The p-value for the interaction term TV × radio is extremely low, indicating that there is strong evidence for $H_A: \beta_3 \neq 0$.
- The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using TV and radio without an interaction term.

Interpretation — continued

- This means that (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.

- The coefficient estimates in the table suggest that an increase in TV advertising of \$1, 000 is associated with increased sales of

$$(\beta_1 + \beta_3 \cdot \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio}$$

- An increase in radio advertising of \$1, 000 will be associated with an increase in sales of

$$(\beta_2 + \beta_3 \cdot \text{TV}) \times 1000 = 19 + 1.1 \times \text{TV}$$

Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not.

- The **hierarchy principle**:
  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.

- The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.

Generalizations

There exist many different extensions and generalizations to the simple linear regression model:

- **Classification problems**: logistic regression, support vector machines
- **Non-linearity**: kernel smoothing, splines and generalized additive models; nearest neighbor methods.
- **Interactions**: Tree-based methods, bagging, random forests and boosting (these also capture non-linearities)
- **Regularized fitting**: Ridge regression and lasso