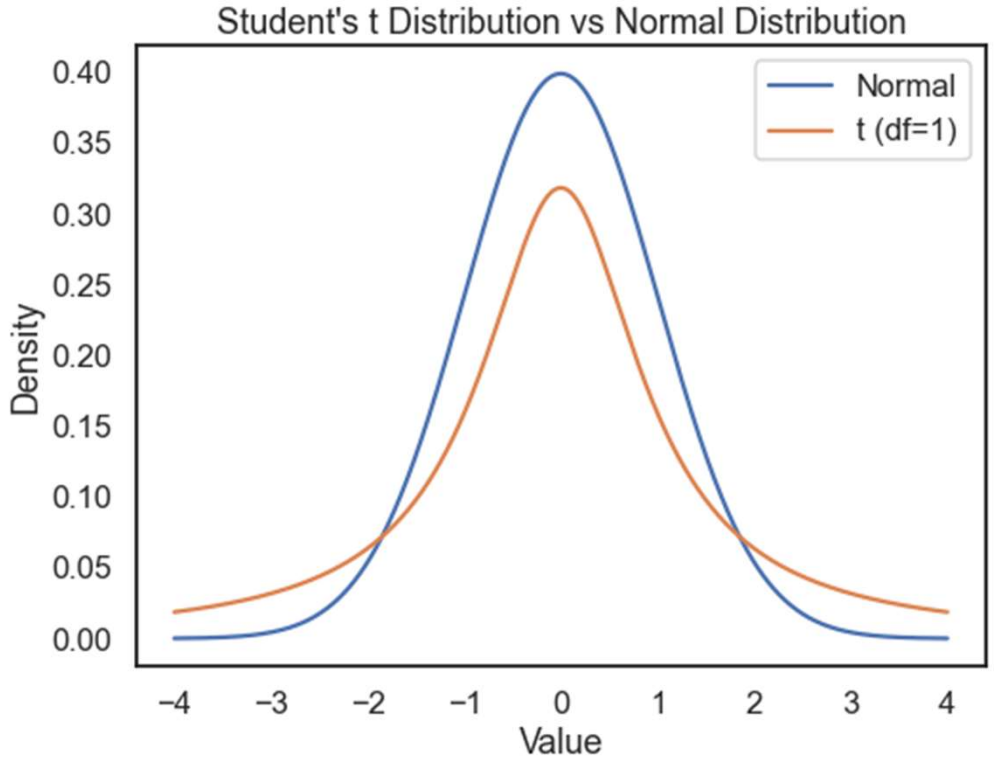
Linear Regression

The Student’s t distribution vs. the Normal distribution

Student’s t distribution by English statistician William Gosset under the pseudonym "Student". Used when the sample size is small (<30)

Normal distribution by Carl Friedrich Gauß



Interval	area
[μ–1σ, μ+1σ]	68.3
[μ–2σ, μ+2σ]	95.4
[μ–3σ, μ+3σ]	99.7

Student t distribution can be applied with small sample size with less than 30

only applicable when having enough samples

Correlation

In statistics, correlation or dependence is any statistical relationship between two random variables (aka. bivariate data), e.g. x_1 and x_2 or x_1 and y, …

The Pearson correlation measure is defined as

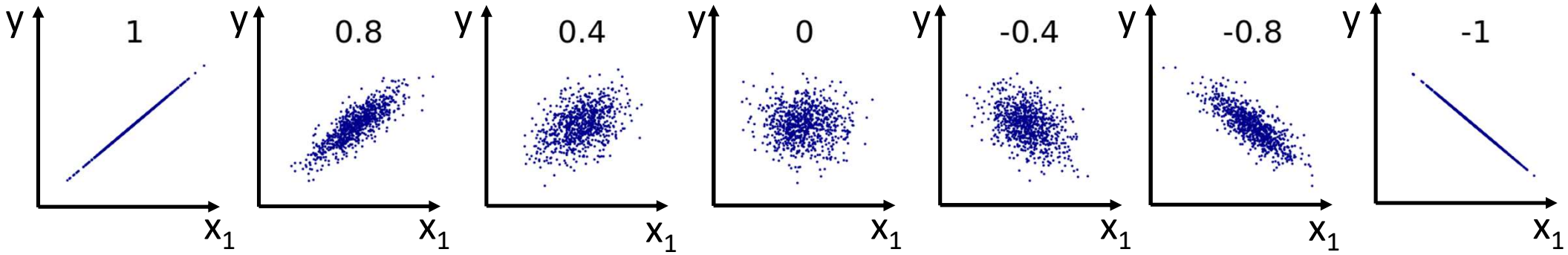
deviation of x to mean of x and y and mean of y

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

covariance

std deviation

Examples for different r:

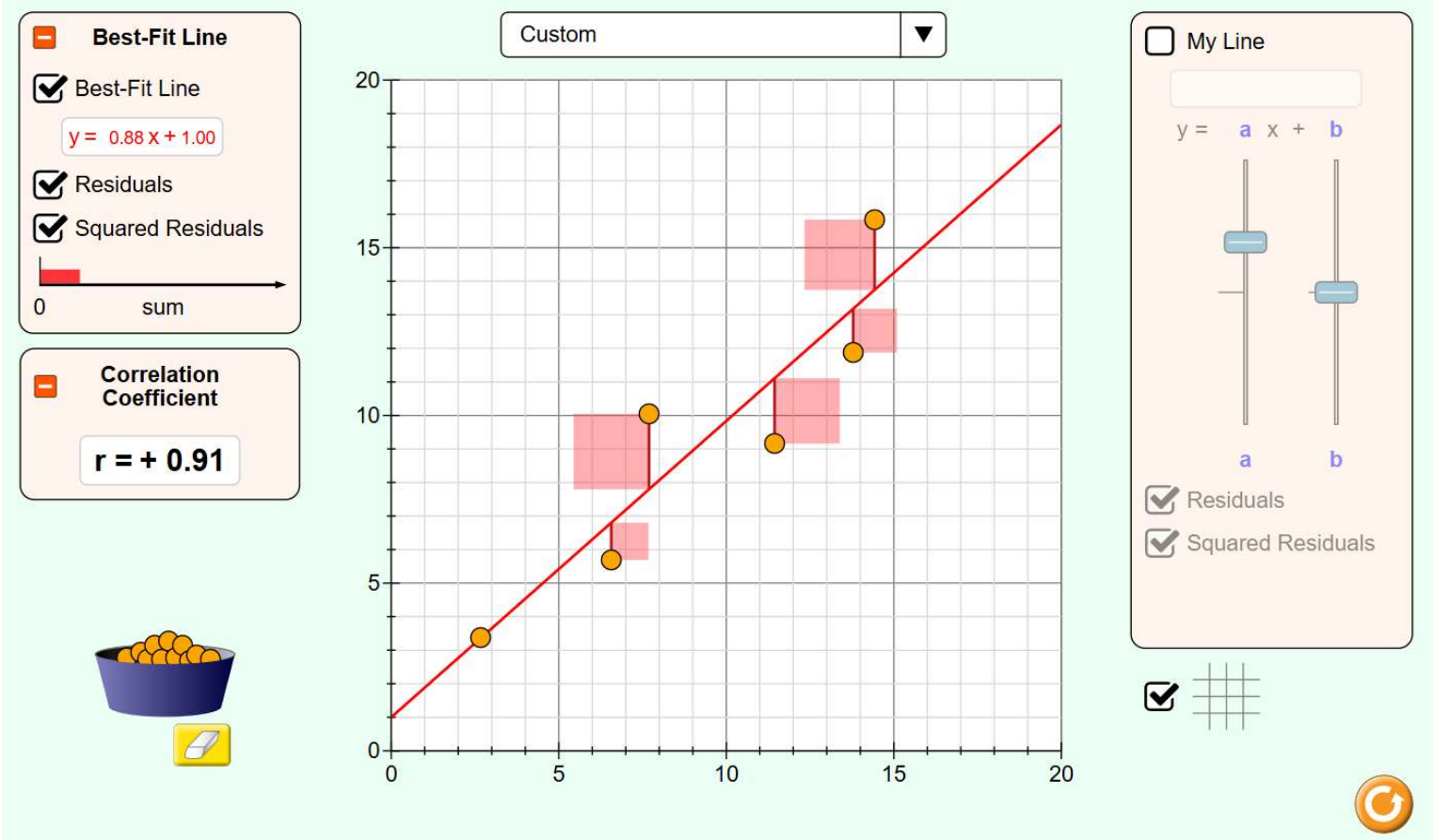


when r = 1, correlatio of x to y is positively dependent

no dependence between x and y

the bigger x, the smaller negatively correlated

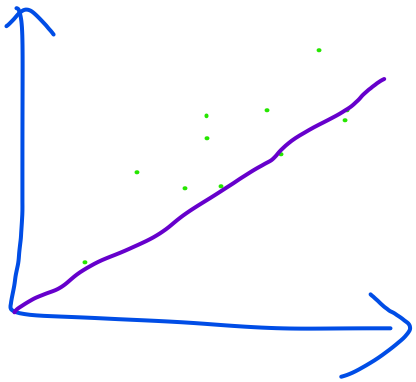
Least squares linear regression - Animation



https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html

Linear Regression

- A simple approach to supervised learning
- It assumes that the dependence of Y on X is linear
- Conceptually and practically very useful
- Which questions we are trying to answer by doing this linear regre
  + is there relationship between certain feature & the response
  + what feature are useful in predicting the response
  + How accurately can we extrapolate
  + The relationship actually linear



y = +b + m.x - linear function

Linear regression using a single predictor X

Let's yi^ = ß0^ + ß1^ xi be the prediction for Y based on the ith value of x with ß0^ as intecept & ß1^as slope. Then e_i = y_i - y_i^ represents the ith residual

residual

We then define the residual sum of squares (RSS) as:
RSS = e_1^2 + e_2^3+...+e_n^2
the least squares approach choose ß0^ and ß1^ to minimize the RSS. the miniimizing values can be shown to be.

ß_1^ = Sum(deviation of x and y)/ Sum (deviation of x)^2 - how to find the slope
ß_0^= sample y - ß_1^(sample x)
=> find y_mean and x_mean (sample means)

Assessing the accuracy of the coefficient estimates

The standard error of an estimate reflects how it varies under repeated sampling.
We have:

Std error can be sued to compute confidence intervals. a 95% confidence interval is defined a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

It has the form: - screenshot

This means, there is aaprox 95% chance that the interval ß1

will contain the true value ß1

Quiz question

Suppose we accidentally doubled our data, leading to observations and error terms identical in pairs. If we ignored this, our standard error calculations would be as if we had a sample of size 2*n*, when in fact we have only *n* samples.

What happens in that case to the confidence interval $\beta_1 \in \left[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)\right]$ with $\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

A) It narrows by a factor of 2

B) It narrows by a factor of $\sqrt{2}$

C) It remains unchanged

D) It widens by a factor of 2

Hypothesis testing

+ std Error can be used to perform hypothesis tests on the coefficients:
 - Null hypothesis H_0: there is no relationship between X and Y, this correspond to H_0: ß_1 =0
 - Alternative hypothesis H_A: there is some relationship between X & Y, this correspond to H_A: ß_1 is different from 0

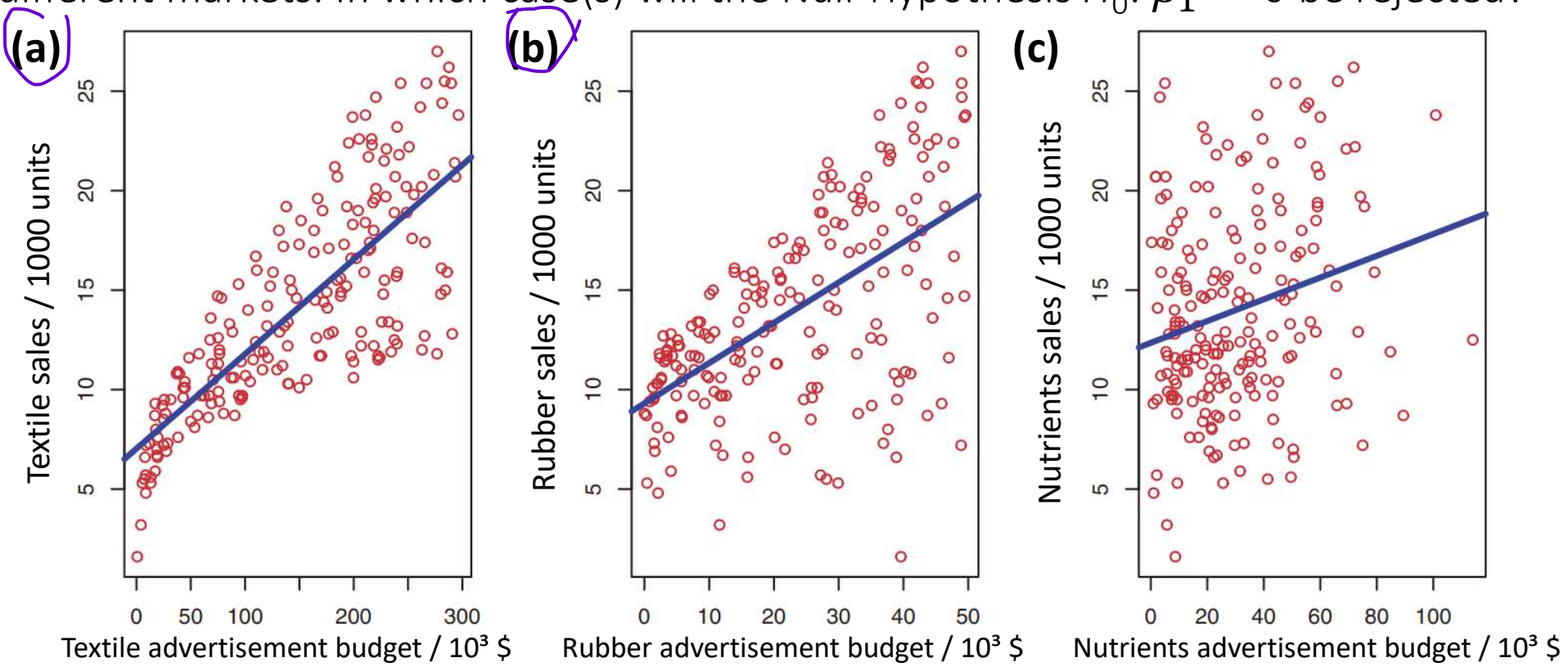
To test the Null hypothesis we compute a t- statistics, given by: (ß1^- ß1)/ SE (ß1^) = ß1^/ SE(ß^1) => t measures the number of stds that ß1^is away from 0

we reject H_0 : ß_1 = 0 when : t is large (usually > 10) & p (the probability of observing any value equal to |t| or larger) value is small (< 0.05)

+ F is large: finds out if at least one predictor is useful F = ((Total sample square - RSS)/p)/ RSS/ (n-p-1)
+ Z is large: measures accuracy estimate by ß^1/ SE(ß1^)

Quiz question

A chemical company spends a certain budget to advertise for textiles, rubber, and nutrients in 200 different markets. In which case(s) will the Null-Hypothesis H_0: $\beta_1 = 0$ be rejected?



p-value:	<0.0001	<0.0001	0.8599
t-statistics:	32.8	21.9	-0.2
Std. error:	0.0014	0.0086	0.0059

p smaller than 0.05 and t greater than 10

Quiz question

The following correlation matrix is provided:

	Textile ad.	Rubber ad.	Nutrients ad.	Sales
Textile ad.	1	0.05	0.06	0.78
Rubber ad.		1	0.35	0.58
Nutrients ad.			1	0.23
Sales				1

Which advertisement seems to have the lowest success in terms of increasing sales?

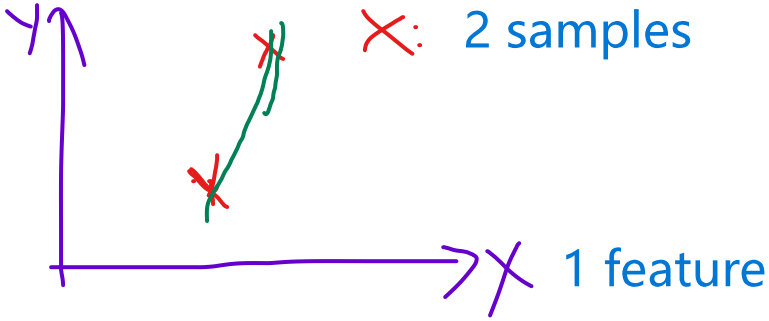
a) Textile ad.

b) Rubber ad.

c) Nutrients ad.

Generalization of least squares linear regression

A supervised parametric model for p (#feature) < n (#sample)



f(x) = ß_o^+ ß_1^x_1 +...+ß_n^x_n + epsilon (irreducible error) with ß_o^as intercept and ß_1^to ß_n^as slope. ß_i^are also called the coefficients or parameters. and epsilon is the irreducible error.

RSS = Residual Sum of square = Sum (i=1..n) (y_i -y mean)^2 with (y_i - y mean) being the ith residual

TAA = Total sum of squares = (Sum (i=1 ... n) (y_i - y_mean))^2
Assessing the overall model accuracy:
R - squared = fraction of variance explained
R^2 = 1 - RSS/TSS

r^2 (correlation) = R^2 in case of linear regression
R^2 belong to [0;1]. The higher R^2 the better is the quality of the prediction

Quiz question

Which of the following are potential problems of linear regression?

a) The response-predictor relationship is non-linear
check the trend in residual plots to find out e_i = y_i - y_i predicted

b) Heteroscedasticity (Non-constant variance in error terms)
For example. time series data

c) Outliers and High-leverage points
outliers y is far from the value predicted by the model
high leverage point uses levarage statistics h

d) Collinearity
splot it by calculating the variance inflation factor

Collinearity

collinearity is the linear relationship between two predictors
- for eg: temp colinear with humidity
- prediction of sickness according to body temp, or temp impacts the state of body sickness

To detect collinearity:
+ look at the correlation matrix of the predictors (if the value is high so the strong correlation is)
+ if collinearity exists between 3 or 4 variables, but not single variables then it is called multi-collinearity: investigate this using variance inflation factor (VIF)
when VIF > 5 signals problems with collinearity

for VIF (ßi^) = 1/ (1-R^2_x|y). R^2 from a regression of X:j onto all of the other predictors

Solutions:
- Drop one of the variables
- Combine the correlated variables into one single predictor. for eg: by averaging them

Adding feature interaction to the model

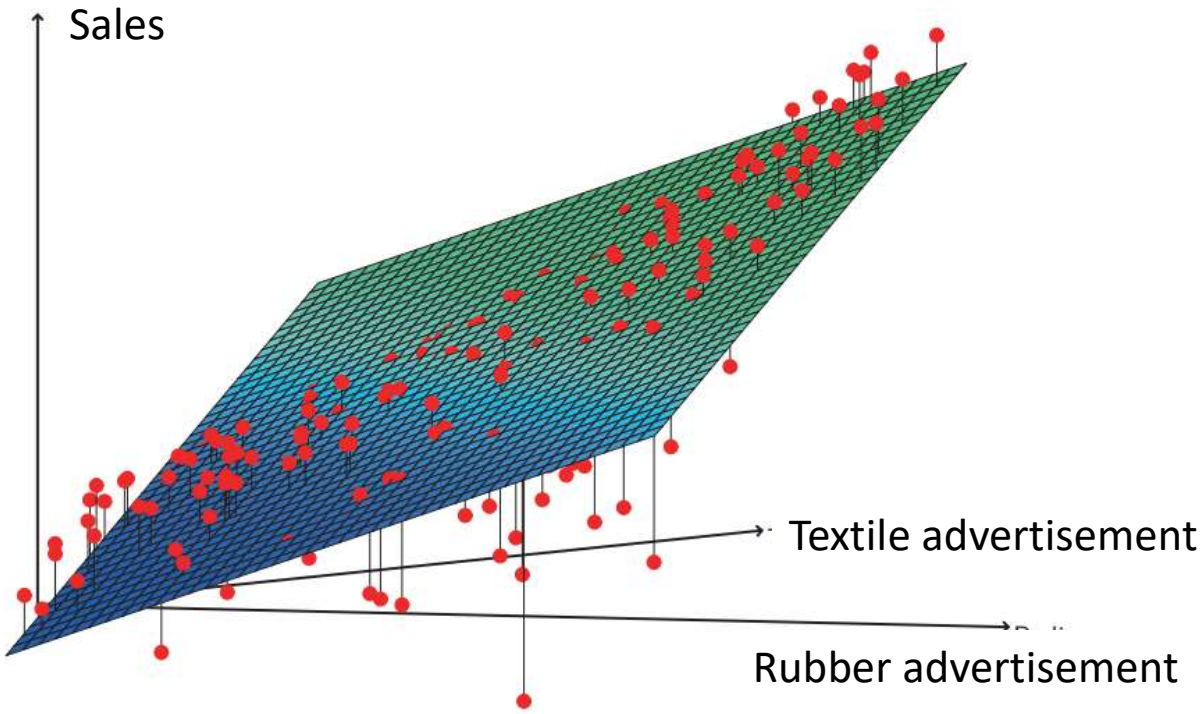
When advertising is split between rubber and textile then the true sales is underestimated. There seems to be a synergy effect, aka. an interaction, between textile and rubber advertisement. Instead of using the model:

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

We add an interaction term:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

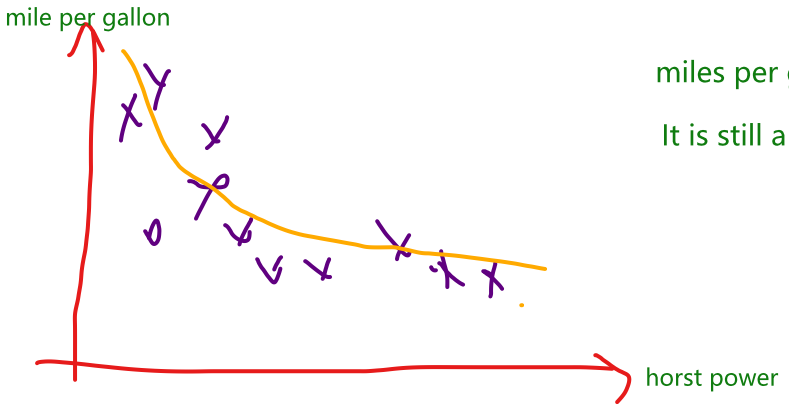
Figure: Introduction to Statistical Learning, T. Hastie et al. 2021



	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.2	$<10^{-4}$
Textile	0.0191	0.002	13.7	$<10^{-4}$
Rubber	0.0289	0.009	3.2	0.001
Textile·Rubber	0.0011	0.001	21.7	$<10^{-4}$

Making the model non-linear

Simple way to add non-linear effects is to: (1) square the features



miles per gallon = ß0 + ß1*horsepwer + ß2.horsepower^2 + epsilon

It is still a linear model, what we have done is x = horse power, x2 = horse power square

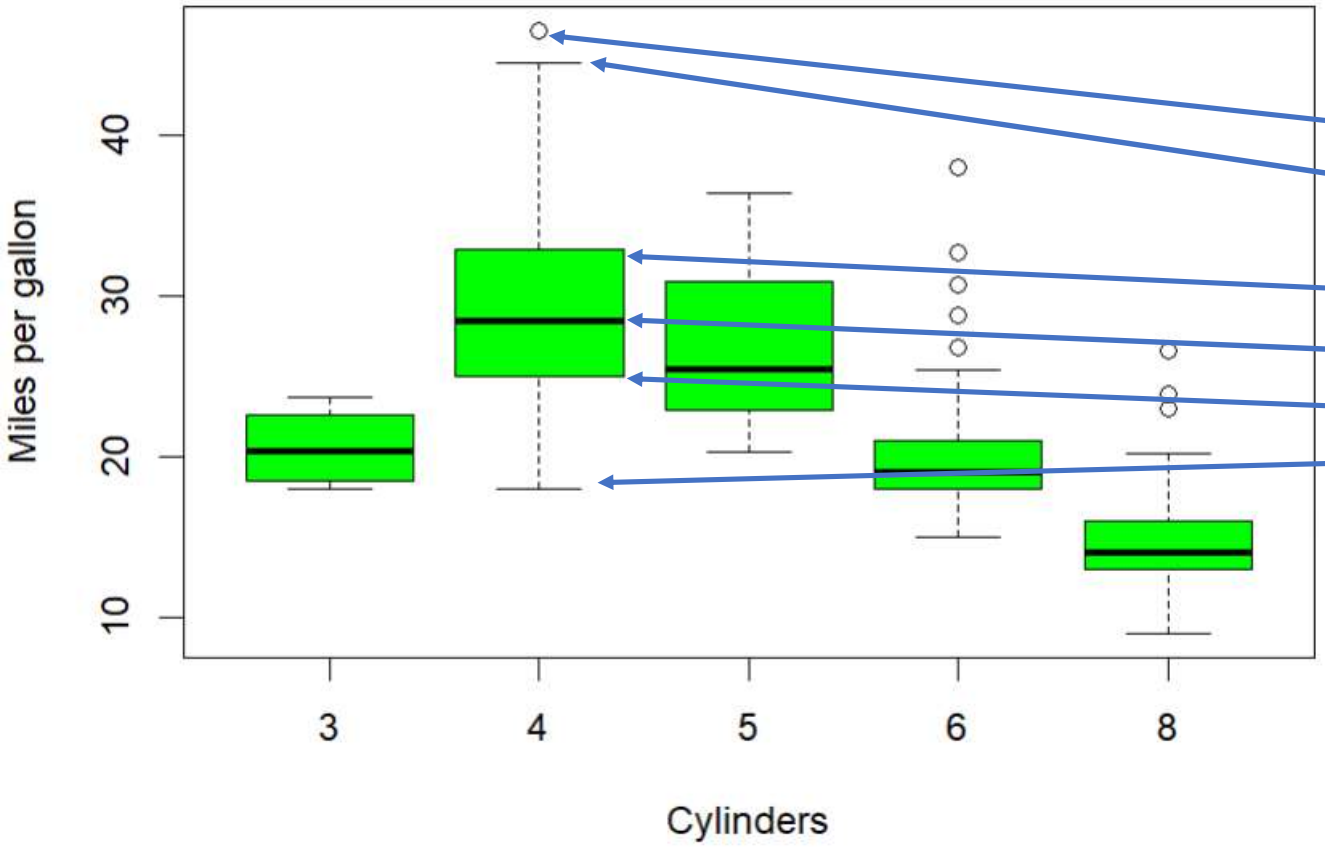
Incorporating qualitative features into the model

example: We want to incorporate into the model whether the car runs on 'diesel, electricity, or Biofuel, ...
For this: we create "dummy variables" an modify the model:

xi_1 = 1: if the car runs on diesel; OR 0 if the car has no diesel engine
x1_2 = 1: if car has electric engine; OR 0 if the car has no electric engine

Therefore, we yi = ß0 + ß_1*xi_1 + ß_2*xi_2 + ... epsilon

Box plots, aka. whisker plots



Box plots:

Outliers above upper whisker

Upper whisker: 3rd Quartile + 1.5 IQR
or upper data point within

Upper box end: 3rd Quartile

Horizontal black line: Median

Lower box end: 1st Quartile

Lower whisker: 1st Quartile - 1.5 IQR
or lowest data point within

Outliers below lower whisker

Box plots for a Normal Gaussian distribution