QBUS6810: Statistical Learning and Data Mining

Lecture 2: Linear Regression

Semester 1, 2019

Discipline of Business Analytics, The University of Sydney Business School

Lecture 2: Linear Regression

1. Introduction
2. The least squares method
3. Interpreting a linear regression model
4. Residual Plots
5. Data transformation
6. Categorical predictors
7. Polynomial regression

Introduction

Linear regression

Linear regression is a simple and widely used method for supervised learning. There are several important reasons for developing an in-depth understanding of this method.

- It is extremely useful conceptually. Many advanced statistical learning methods can be understood as extensions and generalisations of linear regression.
- Due to its simplicity, linear regression is often a good starting point for model building and analysis.
- It works well for prediction in a wide variety of situations.
- It is easy to interpret.

Example: Direct Marketing Data

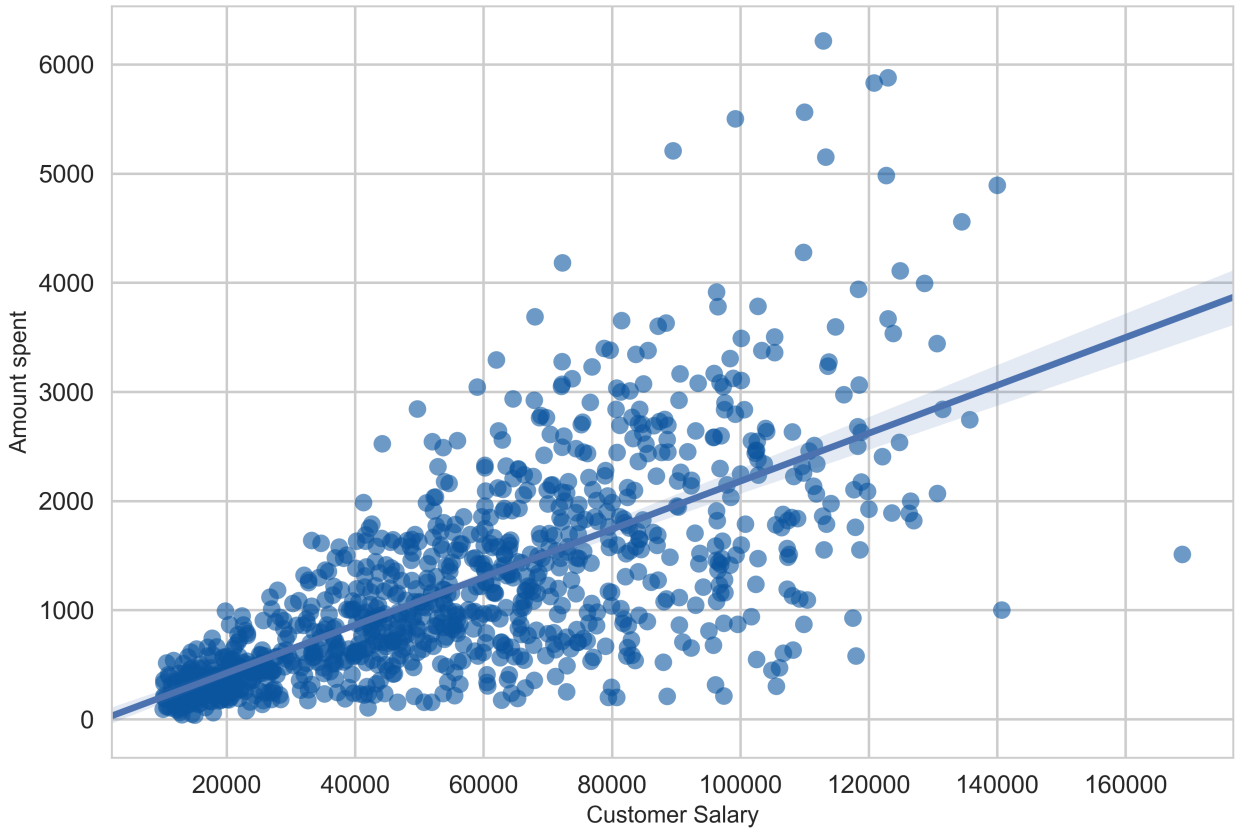
Business problem: Predicting customer purchasing behaviour, with the objective of targeting sales efforts.

Response: Amount spent (dollars) on direct marketing products.

Predictors: Customer salary, age group, gender, marital status, number of children, number of catalogs sent to, spending history (high/low/medium), whether customer is a homeowner, location of nearest physical store (far/close).

The dataset is from Jank (2011).

Exploratory plot



Computing the correlations for the training data reveals that the customer salary is the predictor with strongest linear relationship with the response.

Example: Direct Marketing

Recall the additive error model

$$Y = f(X) + \varepsilon$$

where f is an unknown regression function and ε is random error with the expected value of zero

In the Direct Marketing example we could consider the following model for predicting the amount spent by a customer:

$$\text{AS} = \beta_0 + \beta_1 \times \text{Salary} + \beta_2 \times \text{Catalogs} + \varepsilon$$

The least squares method

Linear regression

In linear regression we assume that

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

leading to the Multiple Linear Regression (MLR) model:

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

Given the values of X, we make the following predictions:

$$\widehat{f}(\boldsymbol{x}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p$$

where coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ are estimated by fitting the model to the training data using the least squares method

Least squares

Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be the training data. For each candidate coefficient vector $\tilde{\boldsymbol{\beta}}$ we define the **residual sum of squares**:

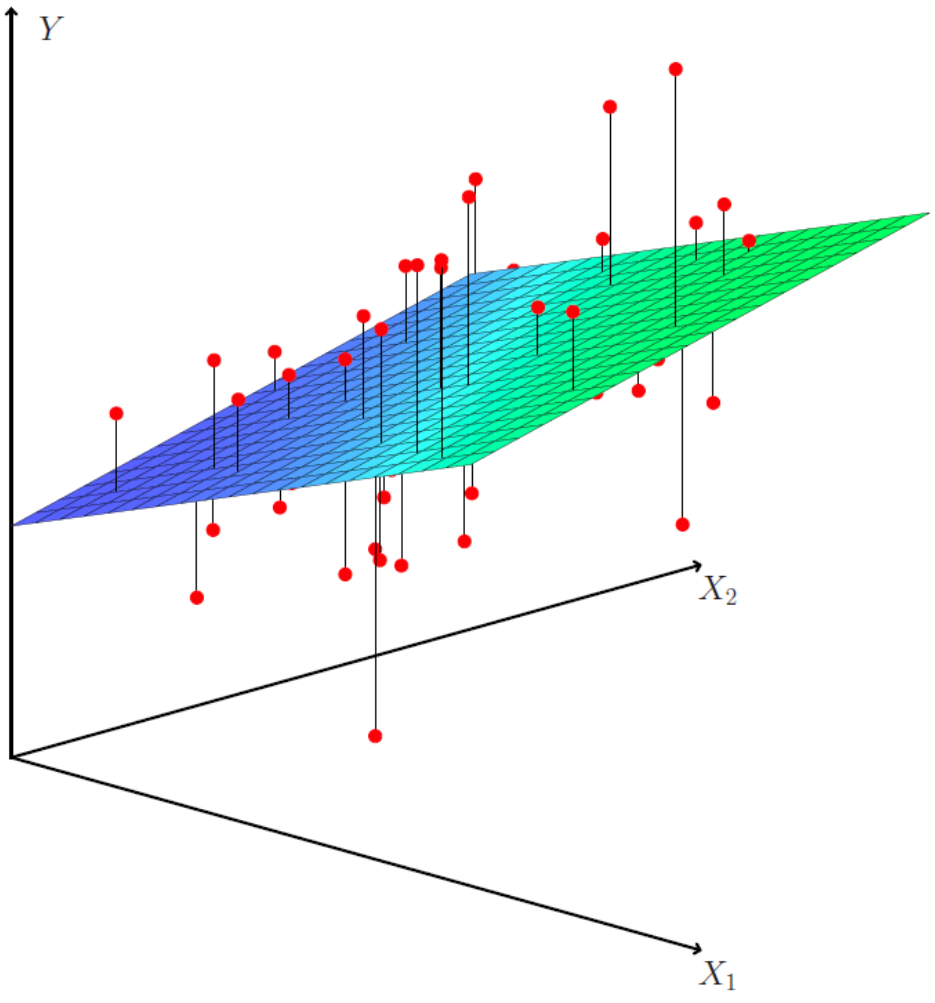
$$\text{RSS}(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \Big(y_i - [\tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \tilde{\beta}_2 x_{i2} + \ldots + \tilde{\beta}_p x_{ip}] \Big)^2$$

最小二乗法

The **ordinary least squares** (OLS) method selects the coefficients that minimise the residual sum of squares:

$$\widehat{\boldsymbol{\beta}} \;=\; \text{minimizer of RSS}$$

Least squares



(Figure from ISL)

Fitted values and residuals

The **fitted values** based on the training inputs are

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_p x_{ip}$$

The regression **residuals** are:

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

Measuring fit

Write $\overline{y}$ for the mean of y_i's. We can show that

y_i = y^_i + e_i

$$\sum (y_i - \overline{y})^2 = \sum (\widehat{y}_i - \overline{y})^2 + \sum e_i^2$$

$$\text{TSS} = \text{RegSS} + \text{RSS}$$

- TSS: total sum of squares.
- RegSS: regression sum of squares.
- RSS: residual sum of squares.

Measuring fit

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

Interpretation: R2测量对应数据中由估计的线性回归模型计算的方差的比例。其数值大小反映了回归贡献的相对程度，即在因归关系所能解释的百分比

- The R^2 measures the proportion of the variation in the response data that is accounted for by the estimated linear regression model.
- The R^2 can not decrease when you add another variable to the model.
- The R^2 is an useful part of the regression toolbox, but (when computed on the training data) does not measure the predictive accuracy of the estimated model. 不测量估计模型的预测准确性。

Interpreting a linear regression model

Example: Direct Marketing

$$\widehat{AS} = -53.68 + 0.0199 \times \text{Salary} + 51.695 \times \text{Catalogs}$$

Interpretation (Salary):

If we select two customers from the population with the same number of Catalogs but a 1 dollar difference in Salary, we estimate that the customer with the higher salary is expected to spend 0.0199 dollars more.

In other words, and in the sense of the sentence above, a 1000 dollar increase in Salary is associated with about a 20 dollar increase in spending.

Interpreting coefficients (mathematical justification)

For example, with $p = 2$ and focusing on the first predictor:

$$\begin{aligned} & E(Y|X_1 = x_1 + 1, X_2 = x_2) - E(Y|X_1 = x_1, X_2 = x_2) \\ &= [\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2] \\ &= \beta_1 \end{aligned}$$

In general:

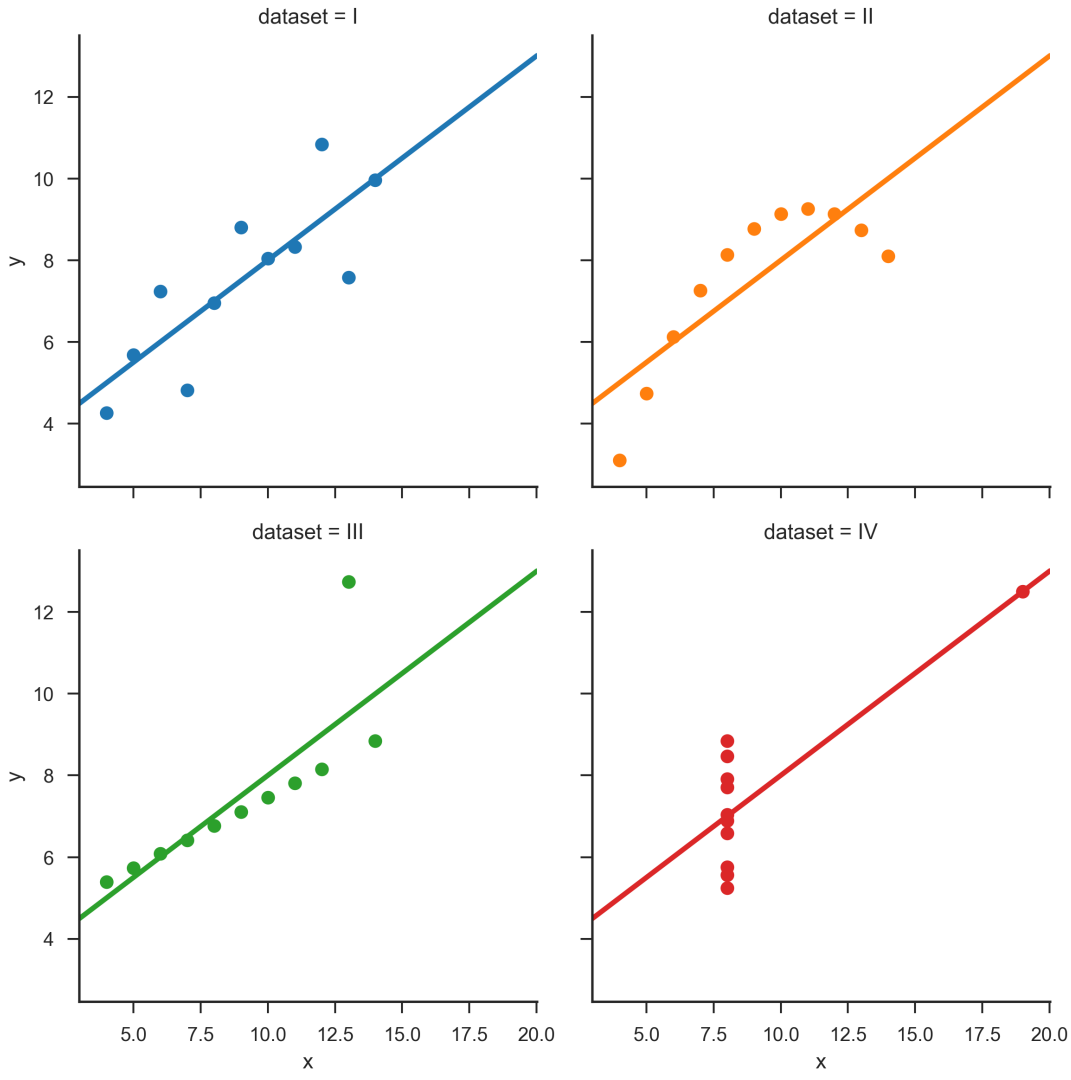
$$\beta_j = E(Y|X_j = x_j + 1, X_{\neq j} = \boldsymbol{x}_{\neq j}) - E(Y|X_j = x_j, X_{\neq j} = \boldsymbol{x}_{\neq j})$$

Residual Plots

Exploratory data analysis

Exploratory data analysis (EDA) is the process of examining and describing data through visualisation and numerical summaries to gain insight, discover structure, and detect potential issues and outliers.

Example: Anscombe Quartet



Example: Anscombe Quartet

In this classic illustrative example by Anscombe (1973):

- In each case, the variables have the same sample mean and variance.
- All the regression lines are the same.
- All the R^2 values are the same.

The figures are the ones telling the story!

Multiple Linear Regression (MLR) model

We can use residual plots to assess the following important assumptions of MLR:

- Linearity: the true regression function is linear, i.e.

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

- Constant error variance: the error term, ε, has variance σ^2, which does not depend on the values of X.

Checking the assumptions

残差在数理统计中是指实际观察值与估计值（拟合值）之间的差。
“残差”蕴含了有关模型基本假设的重要信息。
如果回归模型正确的话，我们可以将残差看作误差的观测值。

It is important to check MLR assumptions with data. The following plots are often useful:

残差

- Fitted values against residuals.
- Fitted values against squared or absolute residuals.

We can complement these plots by using each predictor instead of the fitted values (e.g. plot X_1 vs residuals, X_2 vs residuals, and so on)

Example: Direct Marketing

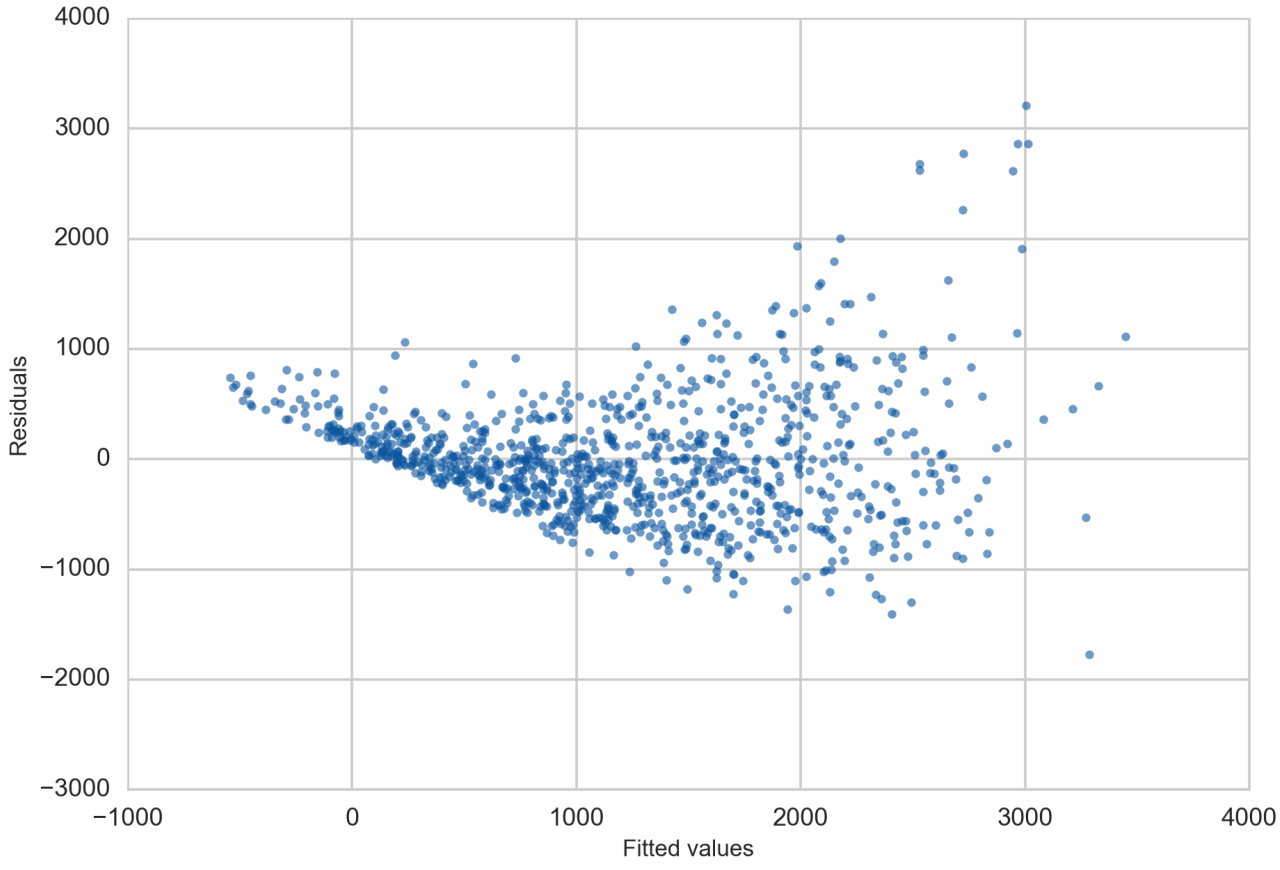
Consider the following estimated model for predicting the amount spent by a customer:

$$\widehat{AS} = -53.68 + 0.0199 \times \text{Salary} + 51.695 \times \text{Catalogs}$$

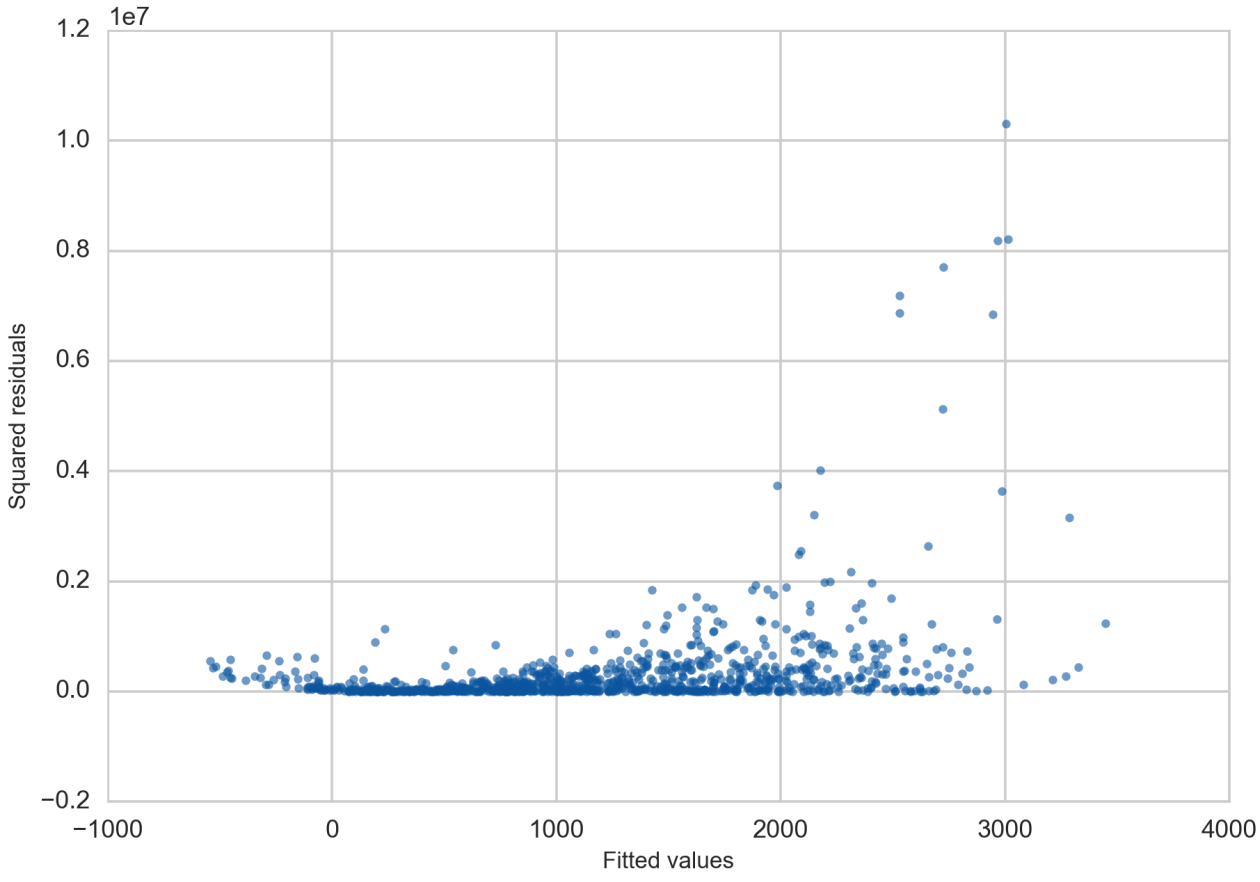
$$R^2 = 0.612$$

We should check whether it is adequate for the data.

Residual plot: fitted values against residuals



Residual plot: fitted values against squared residuals



Example: Direct Marketing

Diagnostics reveal the following problems:

- The residuals follow a nonlinear pattern (suggests violation of the linearity assumption).
  残差遵循非线性模式（暗示违反线性假设）。
  .残差具有非常数方差（暗示违反常数方差假设）。
- The residuals have non-constant variance (suggests violation of the constant variance assumption).

Feature engineering

In machine learning and data science, **feature engineering** is the process of constructing relevant predictors from raw data, particularly through domain knowledge. Often, feature engineering is the main driver of performance improvement.

从原始数据构建相关预测变量，特别是相关领域知识

We will now discuss some tools for feature engineering in linear regression.

Data transformation

Data transformation

Data transformation consists of applying a function to each observation of a response or a predictor. We typically use data transformation with the following goals:

1. Modelling nonlinearity.
2. Meeting the assumption of constant error variance.
3. Improving interpretability.

Log transformation

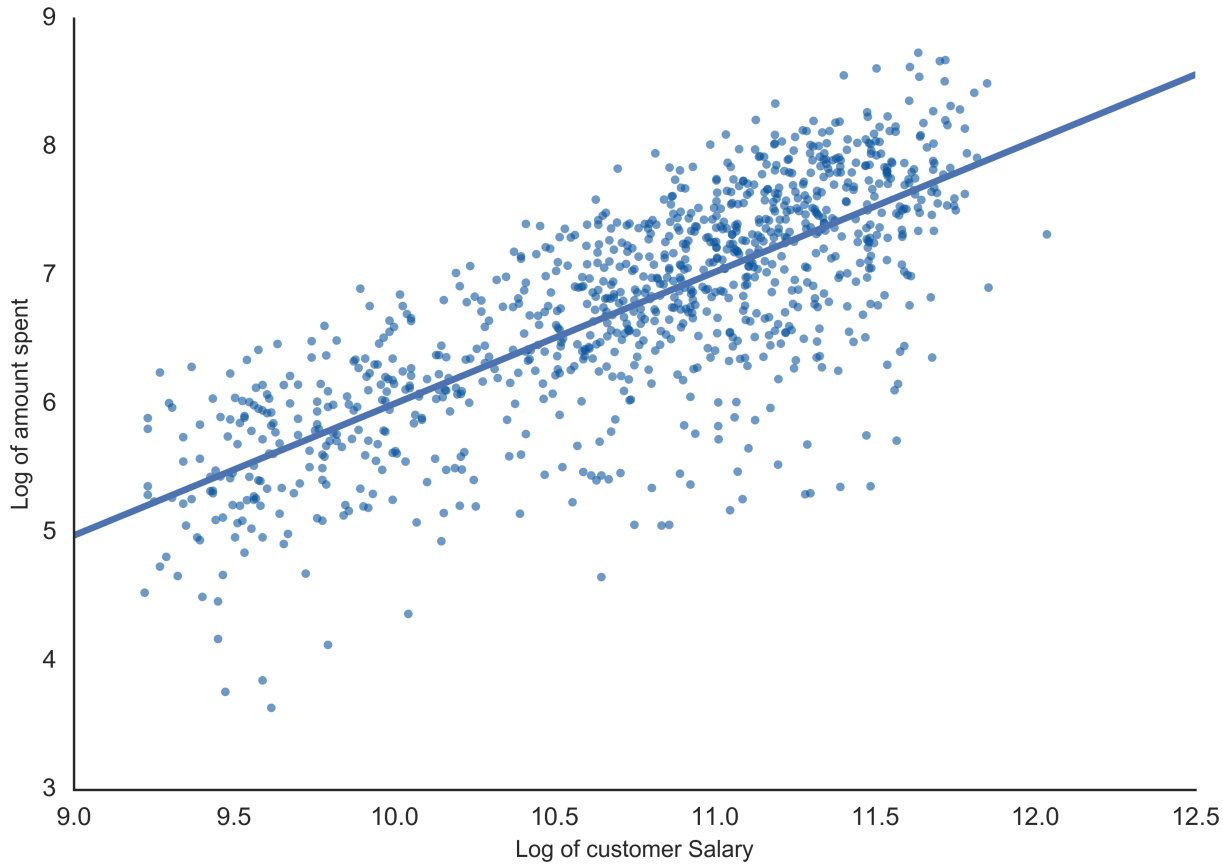
We can improve the model from the last section by considering the following specification:

$$\log(\text{AS}) = \beta_0 + \beta_1 \times \log(\text{Salary}) + \beta_2 \times \text{Catalogs} + \varepsilon,$$

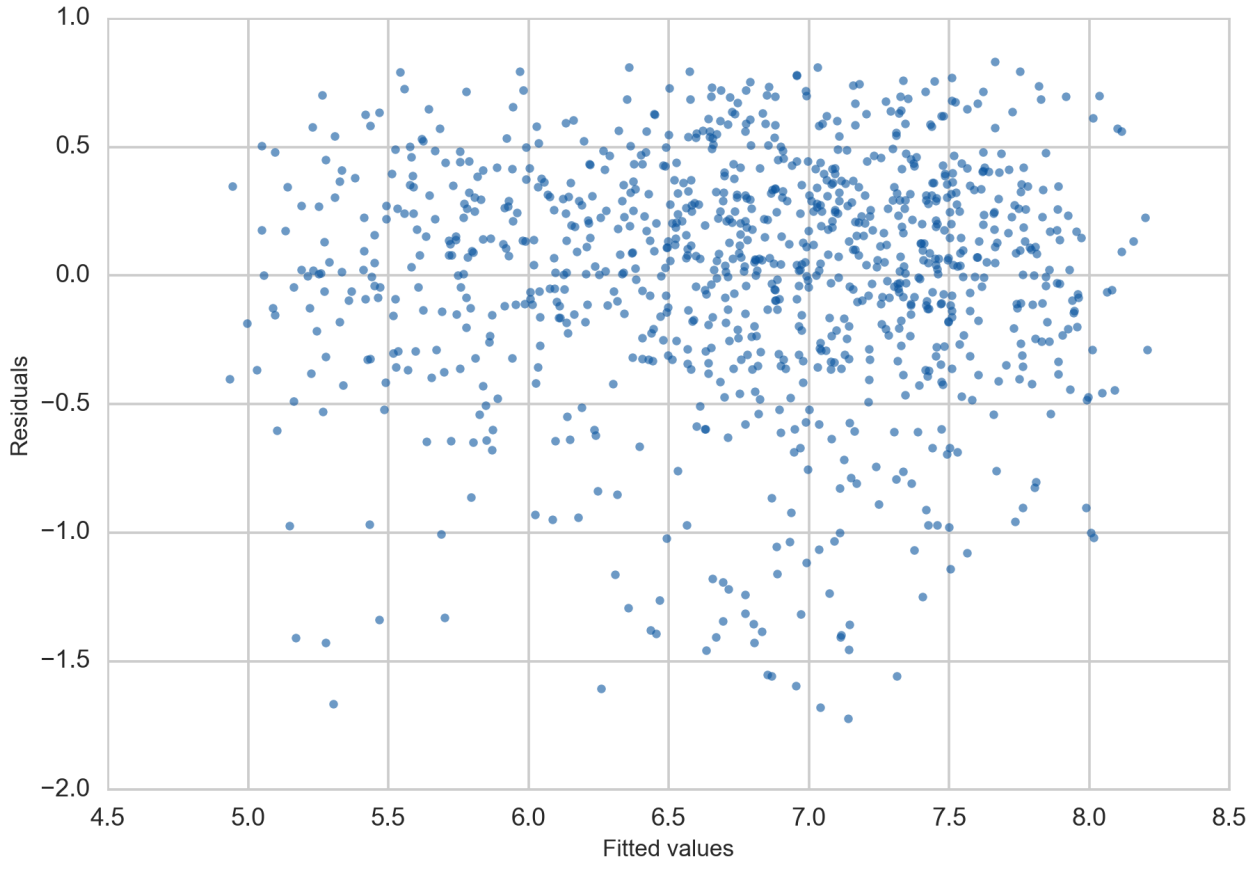
where $\log(\cdot)$ is the natural logarithm. Models with log transformations are particularly relevant for business applications.

We now estimate the new model and look at the diagnostics.

Estimated regression (salary only)



Estimated regression: fitted values against residuals



Example: Direct Marketing

$$\log(\text{AS}) = \beta_0 + \beta_1 \times \log(\text{Salary}) + \beta_2 \times \text{Catalogs} + \varepsilon,$$

Diagnostics reveal that MLR assumptions are much more reasonable in this case, so the new model is preferable.

Questions we need to address:

- What does this model mean? What is the interpretation of β_1?
- How do we predict the amount spent (rather than the transformed variable) with this model?

Log transformation of the response

For concreteness, consider the model with a log transformation of only the response:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon.$$

This model means that for given predictor values,

$$\begin{aligned} Y &= \exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon\right) \\ &= \exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right) \exp(\varepsilon) \end{aligned}$$

Log transformation of the response

After taking the expected value:

$$E(Y|X = \boldsymbol{x}) = \exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right) E(\exp(\varepsilon))$$

We compare this expected value (where $X_1 = x_1$) with the expected value corresponding to $X_1 = x_1 + 1$:

$$\begin{aligned} & E(Y|X_1 = x_1 + 1, X_2 = x_2, ..., X_p = x_p) \\ = \; & \exp(\beta_1) \times E(Y|X_1 = x_1, X_2 = x_2, ..., X_p = x_p). \end{aligned}$$

Interpretation of coefficients

$$
\begin{aligned}
&E(Y|X_1 = x_1 + 1, X_2 = x_2, ..., X_p = x_p) \\
&= \exp(\beta_1) \times E(Y|X_1 = x_1, X_2 = x_2, ..., X_p = x_p).
\end{aligned}
$$

If β_1 is not far from zero,

$$\exp(\beta_1) \approx 1 + \beta_1$$

That is, the expected value is approximately $100 \times \beta_1\%$ larger for $X_1 = x_1 + 1$ than for $X_1 = x_1$.

Interpretation of coefficients log transformations

Case	Regression Specification	Interpretation
Log-Linear	$\log(Y) = \beta_0 + \beta_1 X + \varepsilon$	1 unit increase in X is associated with $\approx 100 \times \beta_1\%$ expected increase in Y.
Log-Log	$\log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon$	1% increase in X is associated with $\approx \beta_1\%$ expected increase in Y.
Linear-Log	$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$	1% increase in X is associated with a $0.01\beta_1$ expected increase in Y.

Estimating the conditional expectation

How can we predict Y given an estimated model

$$\widehat{\log(Y)} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \ldots + \widehat{\beta}_p X_p$$

for $X_1 = x_1, \ldots, X_p = x_p$?

Should we use $\widehat{y} = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots \widehat{\beta}_p x_p\right)$?

It turns out this prediction is downward biased, because 恶化的

$$E(Y|X = \boldsymbol{x}) = \exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right) E(\exp(\varepsilon))$$

and it can be shown that $E(\exp(\varepsilon)) > 1$.

Estimating the conditional expectation

To make a correction, we can use the following modification:

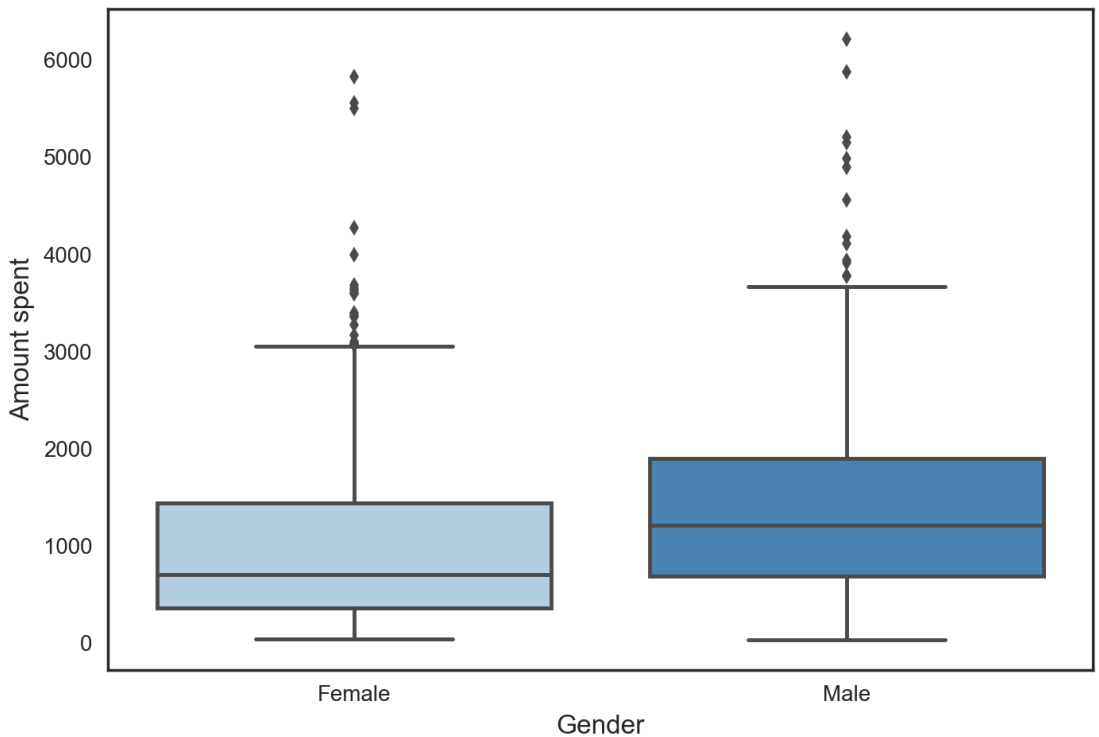
$$\widehat{y} = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots \widehat{\beta}_p x_p\right) \left[(1/n)\sum_{i=1}^{n} \exp(e_i)\right]$$

where e_i is the residual for observation i.

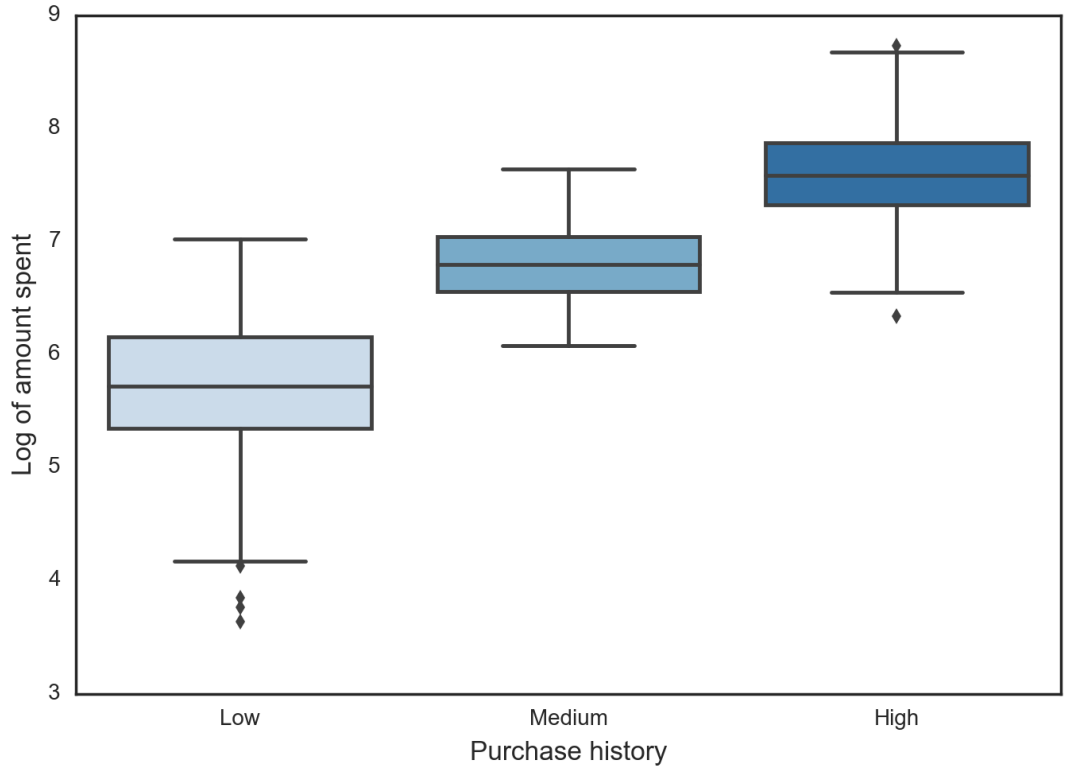
Categorical predictors

Categorical predictors

Several predictors in the direct marketing dataset are categorical. We will now discuss how to incorporate such variables into a regression model.



Exploratory plots



Dummy variables

We start with the simple case of a binary variable: female or male, married or single, whether the stock market goes up or down, etc.

A **dummy variable** codes a binary predictor as a numerical 0 or 1 variable. For example, suppose that we want to construct a predictor to indicate whether the customer is male or female. One option is:

$$X = \begin{cases} 1 & \text{if Male,} \\ 0 & \text{if Female.} \end{cases}$$

Dummy variables

It is good practice to label the dummy variable accordingly:

$$\text{Male} = \begin{cases} 1 & \text{if Male,} \\ 0 & \text{if Female.} \end{cases}$$

The choice of which class label to code as 1 is arbitrary. An equally valid predictor is:

$$\text{Female} = \begin{cases} 1 & \text{if Female,} \\ 0 & \text{if Male.} \end{cases}$$

The two variables above always add up to 1. Hence, we have to choose only one to use as predictor in the regression to avoid redundancy.

Coefficient interpretation

Model:

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

where $D = 0$ or 1 and $X_2, \ldots X_p$ are arbitrary predictors.

任意预测因子。

Interpretation:

β_1 the expected difference in Y when comparing individuals with the same values for all the predictors except D:

$$\begin{aligned}\beta_1 = & E(Y|D = 1, X_2 = x_2, \ldots, X_p = x_p) \\ & - E(Y|D = 0, X_2 = x_2, \ldots, X_p = x_p)\end{aligned}$$

Example: Direct marketing

OLS Regression Results

Dep. Variable:	np.log(AmountSpent)	R-squared:	0.708
Model:	OLS	Adj. R-squared:	0.707
Method:	Least Squares	F-statistic:	803.4
Date:		Prob (F-statistic):	2.48e-265
Time:		Log-Likelihood:	-670.67
No. Observations:	1000	AIC:	1349.
Df Residuals:	996	BIC:	1369.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	-3.7603	0.263	-14.298	0.000	-4.276	-3.244
np.log(Salary)	0.9174	0.024	37.684	0.000	0.870	0.965
Catalogs	0.0475	0.002	20.559	0.000	0.043	0.052
Female	-0.0575	0.031	-1.842	0.066	-0.119	0.004

Example: Direct marketing

$$\widehat{\log(\text{AS})} = -3.76 + 0.917 \times \log(\text{Salary}) + 0.048 \times \text{Catalogs} - 0.058 \times \text{Female}$$

Interpretation:

If we compare a male and a female customer with the same salary and number of catalogs sent, we estimate that the female customer is expected to spend about 6% less.

Binary categories

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

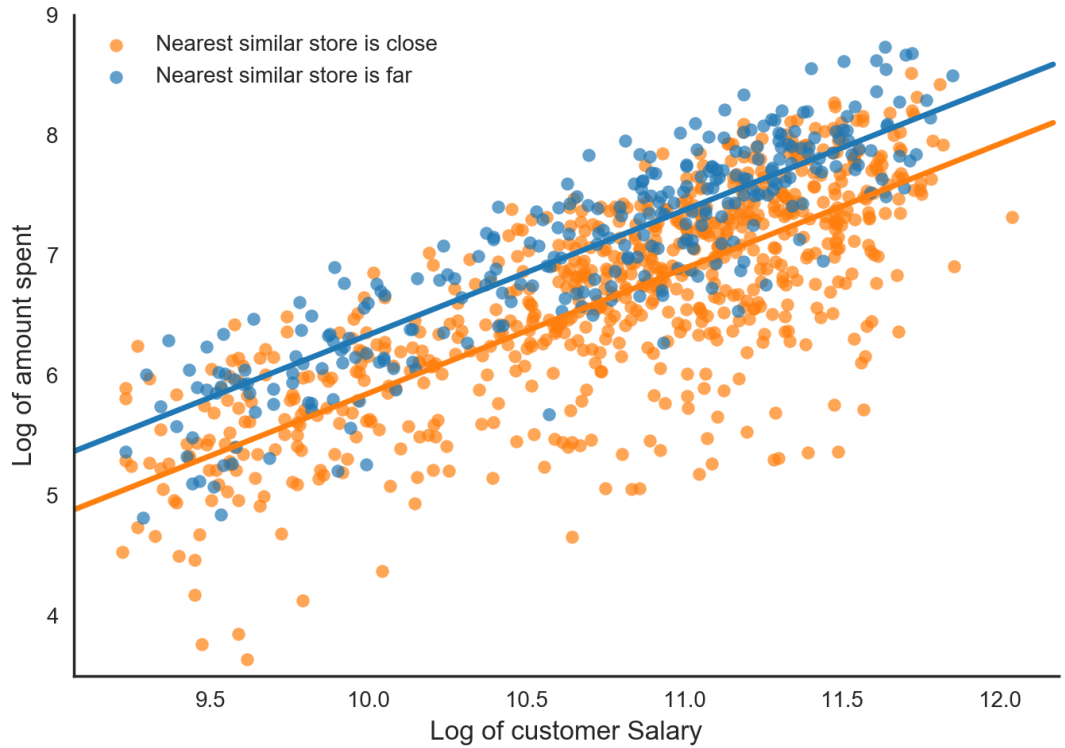
Another way to interpret the model is to think of it as specifying different intercepts depending on the class.

$$Y = \begin{cases} \beta_0 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon & \text{for } D = 0 \\ (\beta_0 + \beta_1) + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon & \text{for } D = 1 \end{cases}$$

Example: Direct Marketing

Consider the model:

$$\log(\text{AS}) = \beta_0 + \beta_1 \times \log(\text{Salary}) + \beta_2 \times \text{Close} + \varepsilon$$



The fitted lines are parallel.

Multiple categories

- Consider now a categorical variable with k classes.
- For this general case, we choose one class to be the baseline and use $k-1$ dummy variables to code the categorical variable.

Multiple categories

Consider the purchase history variable in the direct marketing data. This is a categorical variable with $k = 3$ possible values: $\{\text{High}, \text{Medium}, \text{Low}\}$.

We need to create 2 dummy variables. For example, we can choose Low as the baseline case and define:

$$\text{Medium} = \begin{cases} 1 & \text{if Medium,} \\ 0 & \text{Otherwise.} \end{cases}$$

$$\text{High} = \begin{cases} 1 & \text{if High,} \\ 0 & \text{Otherwise.} \end{cases}$$

General formulation

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_{k-1} D_{k-1} + \beta_k X_k + \ldots + \beta_p X_p + \varepsilon$$

Interpretation:

The regression coefficient β_1 is the expected difference in Y between category one ($C = 1$) and the baseline category (k), conditional on the values of the other predictors:

$$\begin{aligned} \beta_1 = & E(Y|C = 1, X_k = x_k, \ldots, X_p = x_p) \\ & - E(Y|C = k, X_k = x_k, \ldots, X_p = x_p) \end{aligned}$$

Interpretation for the rest of β_j is analogous

Multiple categories

$$\widehat{\log(\text{AS})} = -2.91 + 0.834 \times \log(\text{Salary}) + 0.044 \times \text{Catalogs} + 0.01 \times \text{Medium} + 0.314 \times \text{High}$$

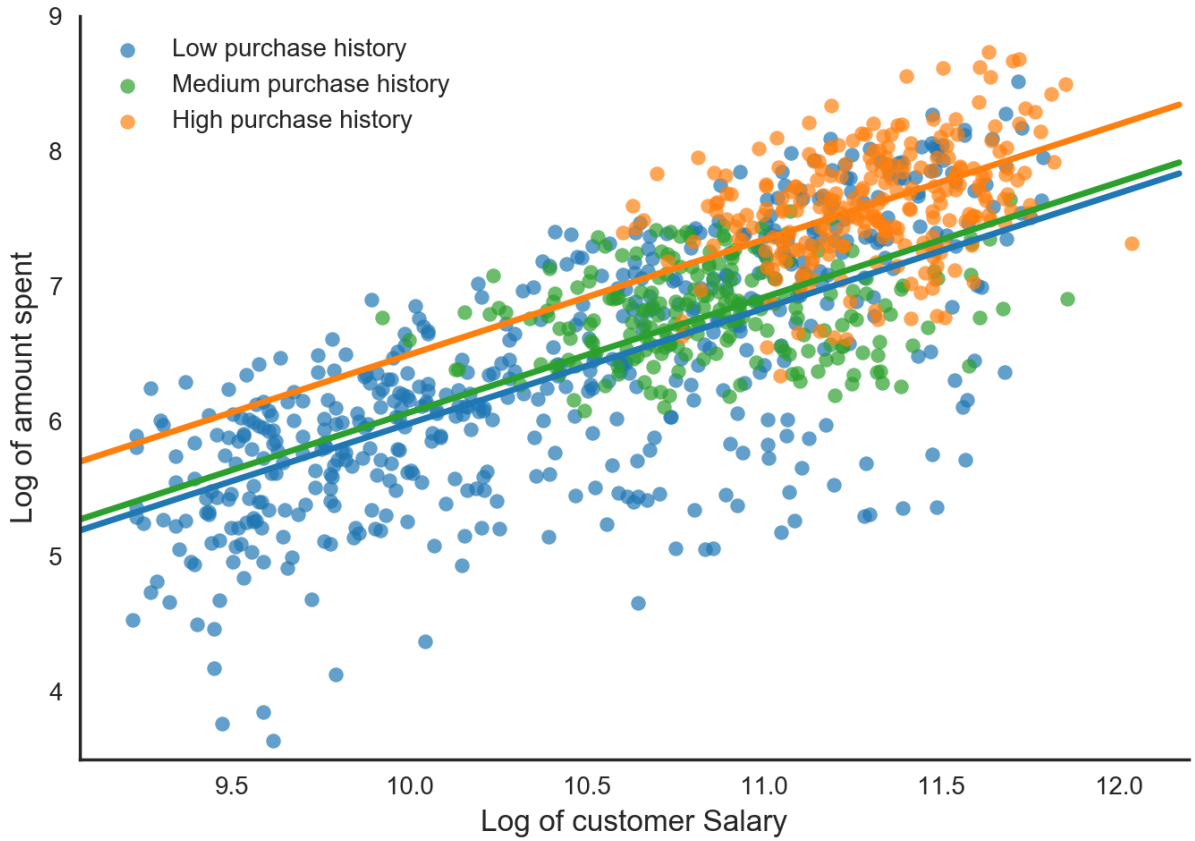
Interpretation:

If we select two customers with the same salary and number of catalogs sent, but one has a Medium purchase history and the other Low, we expect the customer with the Medium purchase history to spend about 1% more.

Example: Direct Marketing

Consider the model:

$$\log(\text{AS}) = \beta_0 + \beta_1 \times \log(\text{Salary}) + \beta_2 \times \text{Medium} + \beta_3 \times \text{High} + \varepsilon$$



Interaction modelling

Suppose that we have a categorical predictor C and a quantitative predictor X_j. In interaction modelling we allow the relationship between Y and X_j to change depending on the category of C.

We achieve this by including as predictor variables in the model the products of the dummy variables for C and the predictor X_j.

For example, suppose that C is coded using dummy variables D_1,...,D_{k-1}. To model the interaction between C and X_j, we include products $(D_1 \times X_j)$, $(D_2 \times X_j)$,...,$(D_{k-1} \times X_j)$ as predictor variables in the model.

Interactions

For example, in the case when C has two categories (one dummy):

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (D \times X) + \varepsilon$$

Three categories (two dummies):

$$Y = \beta_0 + \beta_1 X + \beta_2 D_1 + \beta_3 D_2 + \beta_4 (D_1 \times X) + \beta_5 (D_2 \times X) + \varepsilon$$

Example: Direct Marketing

$$\widehat{\log(\text{AS})} = -3.06 + 0.848 \times \log(\text{Salary}) + 0.044 \times \text{Catalogs} + 2.86 \times \text{High} - 0.226 \times \text{High} \times \log(\text{Salary})$$

This is a regression with different intercepts and salary slopes conditional on the purchase history category:

$$\widehat{\log(\text{AS})} = \begin{cases} -0.20 + 0.622 \times \log(\text{Salary}) + 0.044 \times \text{Catalogs} & \text{if High} \\ -3.06 + 0.848 \times \log(\text{Salary}) + 0.044 \times \text{Catalogs} & \text{if not} \end{cases}$$

The interaction coefficient, -0.226, is the difference in the slopes for $\log(\text{Salary})$

Interactions

Consider the model: $\log(\text{AS}) = \beta_0 + \beta_1 \times \log(\text{Salary}) + \beta_2 \times \text{High} + \beta_3 \times \text{High} \times \log(\text{Salary}) + \varepsilon$

The fitted lines are not parallel.

Polynomial regression

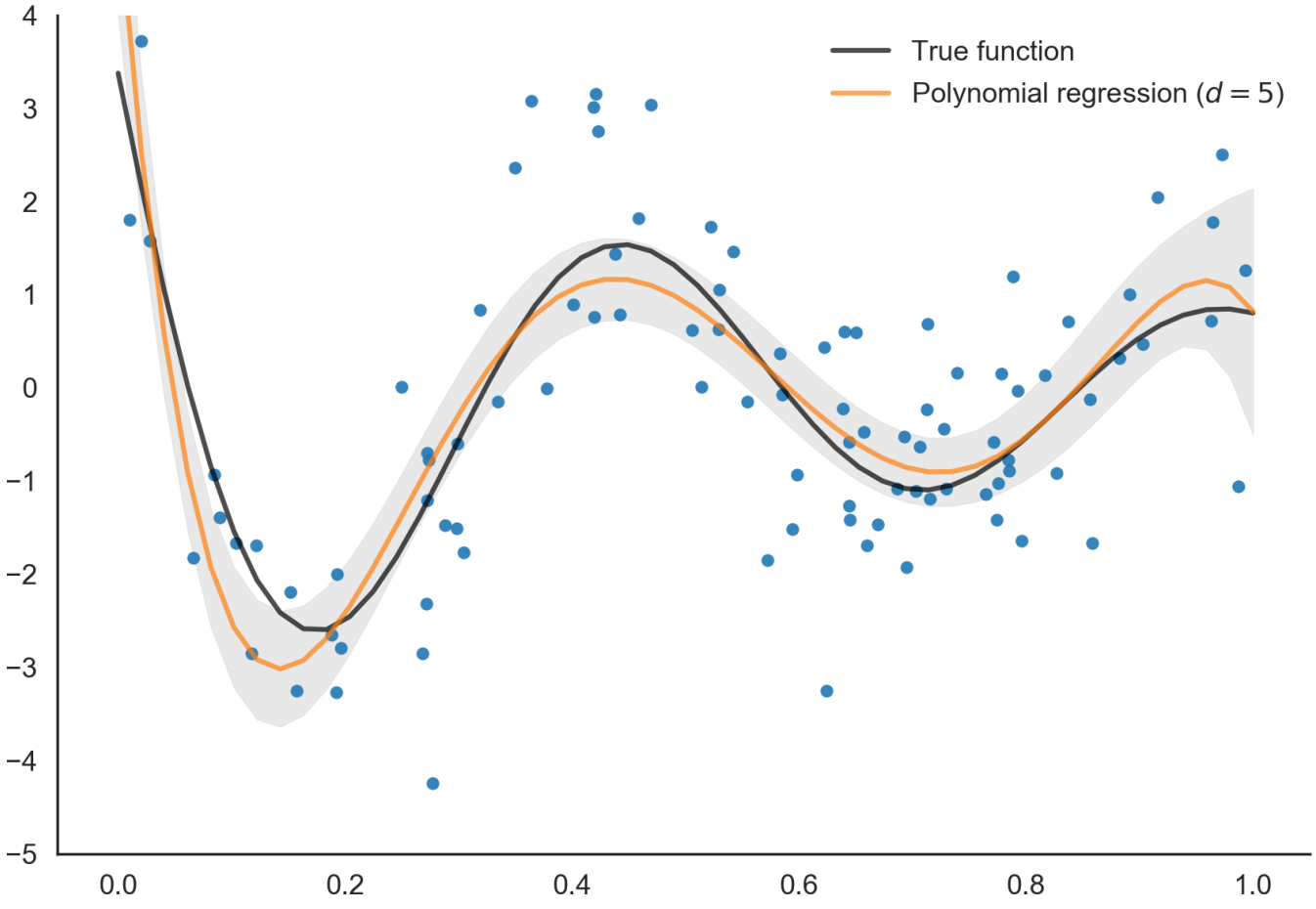
Polynomial regression

The **polynomial regression model** allows us to model a nonlinear relationship between the response and a predictor. The model equation is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_d X^d + \varepsilon,$$

where d is the called the degree of the polynomial. This is an MLR model and all the same methods apply.

Polynomial regression: illustration



Why polynomials?

- Polynomials are mathematically simple.

- A polynomial of sufficiently high degree can approximate any smooth function $f(\cdot)$ arbitrarily well within a certain interval.

有足够高度的多项式可以在一定间隔内任意地近似任何平滑函数f（·）。

Limitations

- Polynomials can overfit. A polynomial of degree d can fit $d + 1$ points exactly, so increasing d produces a wiggly curve that gets close to the data, but predicts poorly.

- Polynomials display a highly nonlocal fit: observations in one region, especially outliers, can seriously affect the fit in another region.

- Polynomials are unstable near the boundaries of the data.

Bivariate polynomials 二元多项式

Suppose that we believe that the data follows a model of the form

$$Y = \beta_0 + f(X_1, X_2) + \varepsilon,$$

where $f(X_1, X_2)$ in unknown.

One option in this case is to use bivariate polynomial regression as approximation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

A special case is

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

which is often presented as the standard interaction model.

Some review questions

- How do we obtain the OLS estimates?
- What is the interpretation of a linear regression coefficient?
- What are some reasons for doing data transformations?
- What is the interpretation of regression coefficients with log transformations?
- How do we compute predictions for a model in which we applied a log transformation to the response?
- How do we include categorical predictors in a regression?
- What is interaction modelling?
- What is polynomial regression and why do we use it?