Statistical Learning and Data Mining

Module 4: Regression Modelling

Semester 2, 2017

Discipline of Business Analytics, The University of Sydney Business School

Module 4: Regression Modelling

1. Introduction
2. The importance of EDA
3. Diagnostics (worked example)
4. Data transformation
5. Categorical predictors
6. Interactions
7. Polynomial regression

Introduction

Regression modelling

In this module we discuss several useful tools for building linear regression models for supervised learning. These tools will also be relevant for more advanced methods.

To complement this content, you should also read the discussion of potential problems with linear regression in Chapter 3 of ISL.

Regression modelling: study notes

Note that the lecture does not have time to cover everything, as this is not a unit on regression analysis. You are expected to study the rest on your own, and read the textbook in detail to develop a solid understanding of the material.

I designed the slides to provide a more comprehensive presentation than the lecture for your benefit and reference. This is all useful knowledge, typically presumed for statistical learning.

Regression modelling

We discuss three related aspects of modelling.

Exploratory data analysis (EDA). Start by understanding the data.

Model building. Here we focus on response transformations and constructing predictors.

Diagnostics. Checking whether the assumptions of a model are valid for the data allows us specify better models, select appropriate methods, and know the limitations of the analysis.

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.

> *Coming up with features is difficult, time-consuming, requires expert knowledge. Applied machine learning is basically feature engineering.* – Andrew Ng

This module discusses tools for feature engineering.

Case study: Direct Marketing Data

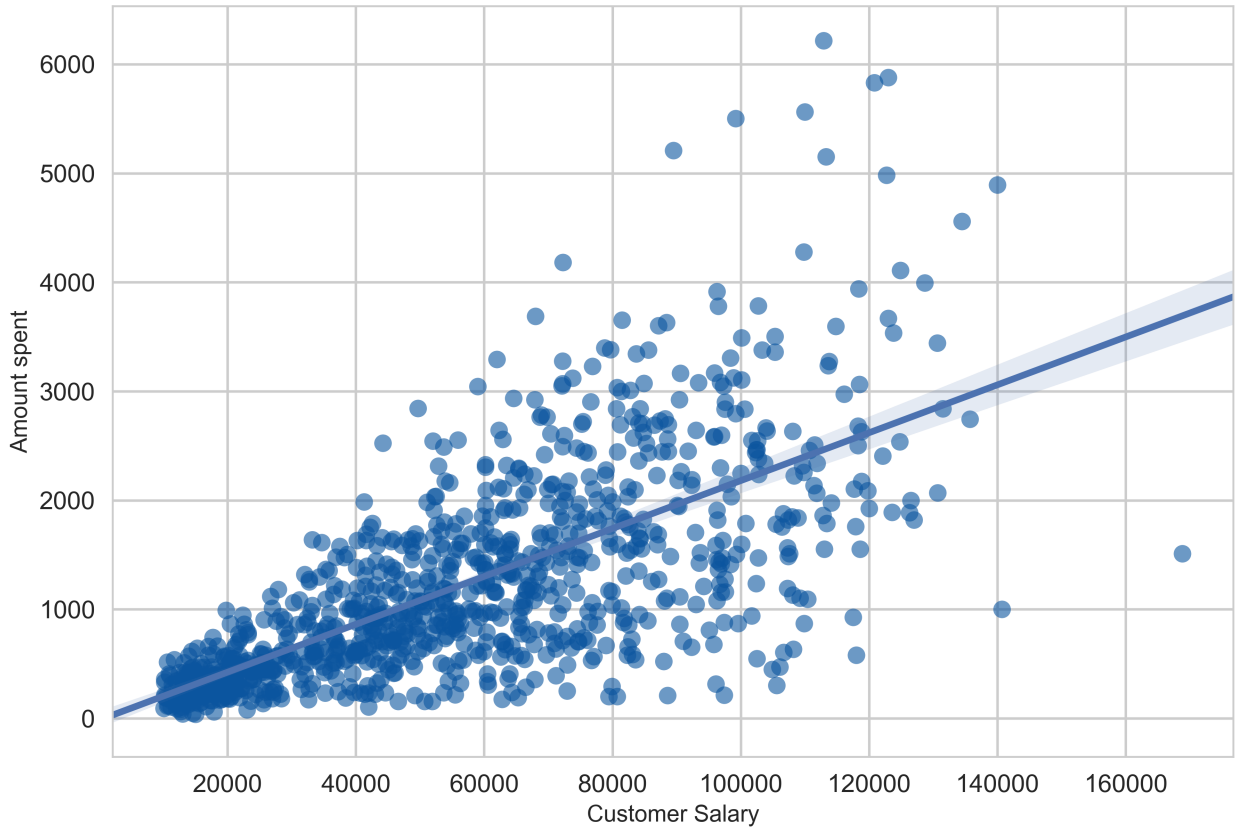
Business problem: Predicting customer purchasing behaviour, with the objective of targeting sales efforts.

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.

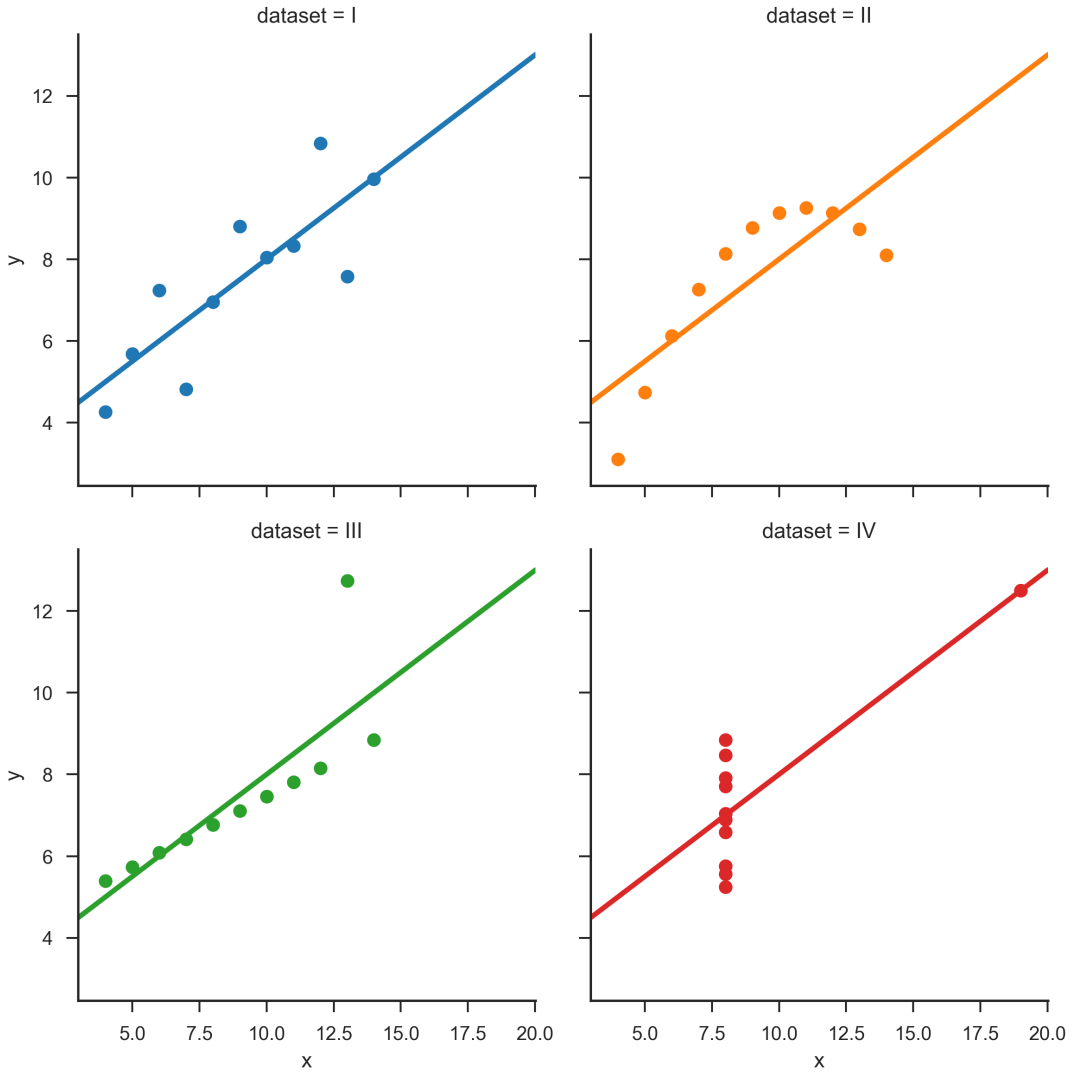
The importance of EDA

Exploratory data analysis (key concept)

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.

Basic statistics review. You should be familiar with plots such as histograms, box plots, scatter plots, quantile-quantile plots, and kernel density estimate (KDE) plots. You should know how to use numerical summaries to describe the center, spread, and shape of a distribution. You be familiar with the concept of the skewness of a distribution.

Example: Anscombe Quartet



Example: Anscombe Quartet

In this classic illustrative example by Anscombe (1973):

- The response and predictor have the same sample average and variance.
- The regression lines are the same.
- The R^2 is the same. Therefore the sample correlations are the same.

The figures are the ones telling the story!

Diagnostics (worked example)

Example: Direct Marketing

We consider the following specification towards building a predictive model for the amount spent by a customer:

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

Example: Direct Marketing

$$\widehat{AS} = \underset{(659.15)}{-53.68} + \underset{(0.001)}{0.0199} \times \text{Salary} + \underset{(2.912)}{51.695} \times \text{Catalogs}$$

$$R^2 = 0.612,\ \overline{R}^2 = 0.611$$

After the estimating the model, we should check whether it is adequate for the data.

Review: MLR model assumptions

1. Linearity: if $X = \boldsymbol{x}$, then

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

   for some population parameters $\beta_0, \beta_1, \ldots, \beta_p$ and a random error ε.

2. The conditional mean of ε given X is zero, $E(\varepsilon|X) = 0$.

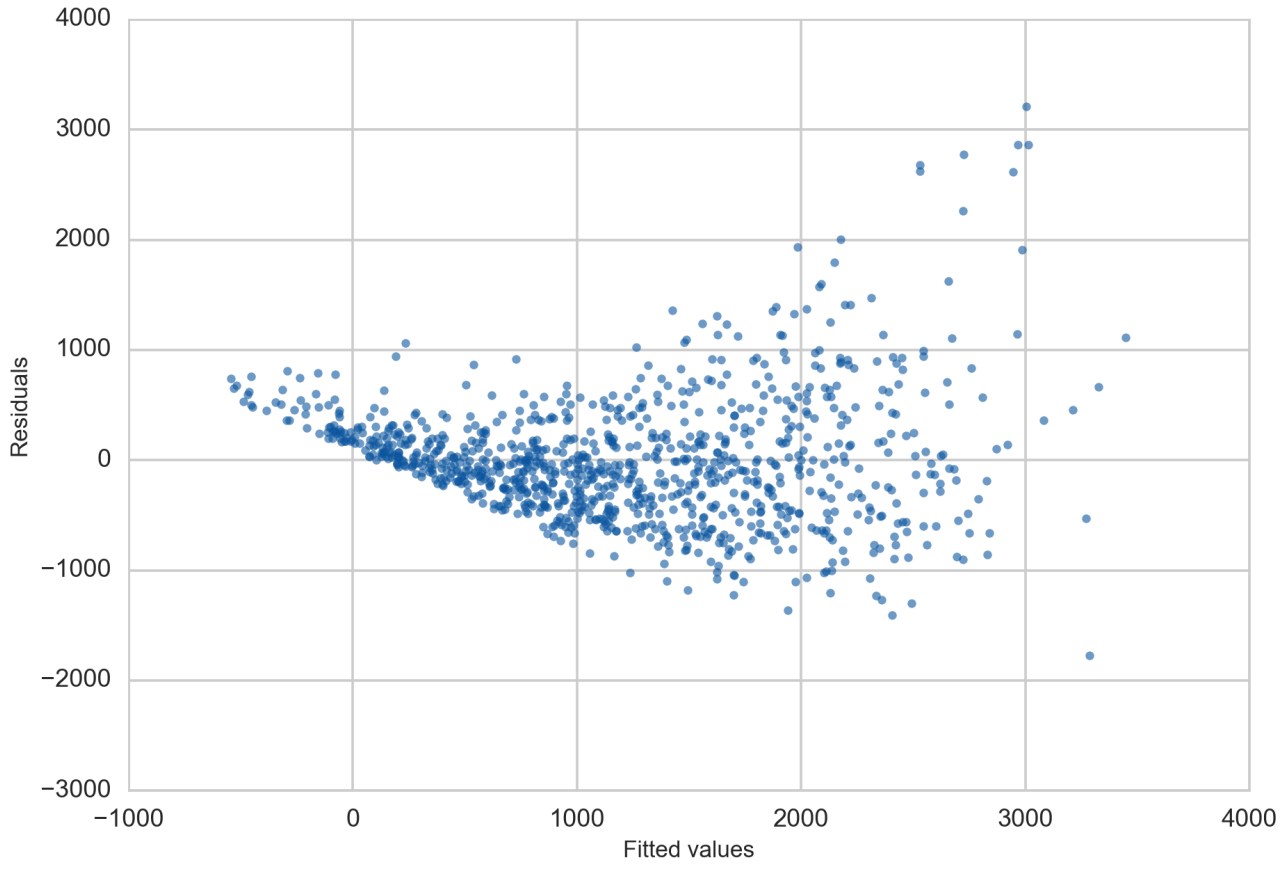
3. Constant error variance: $\text{Var}(\varepsilon|X) = \sigma^2$.

4. Independence: the observations are independent.

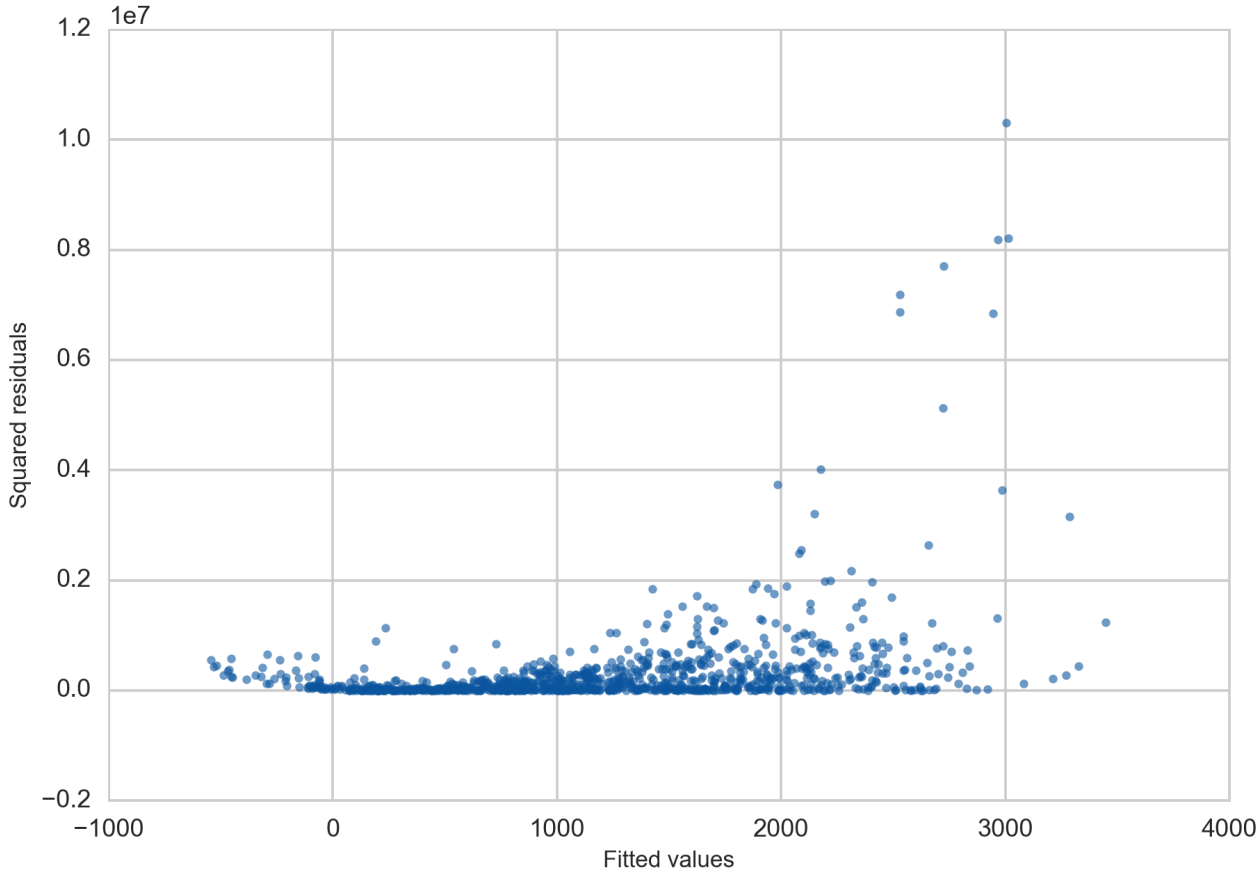
5. The distribution of $X_1, \ldots, X_p$ is arbitrary.

6. There is no perfect multicollinearity (no column of $\boldsymbol{X}$ is a linear combination of other columns).

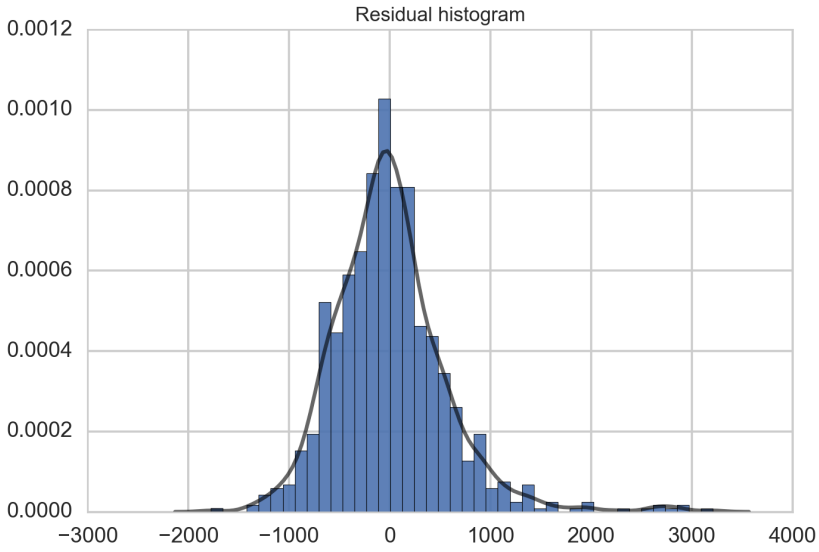
Assumption checking: fitted values against residuals

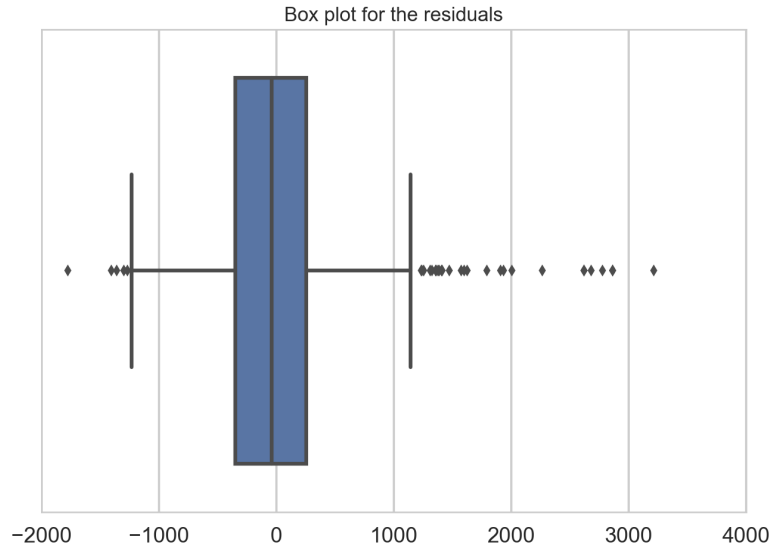


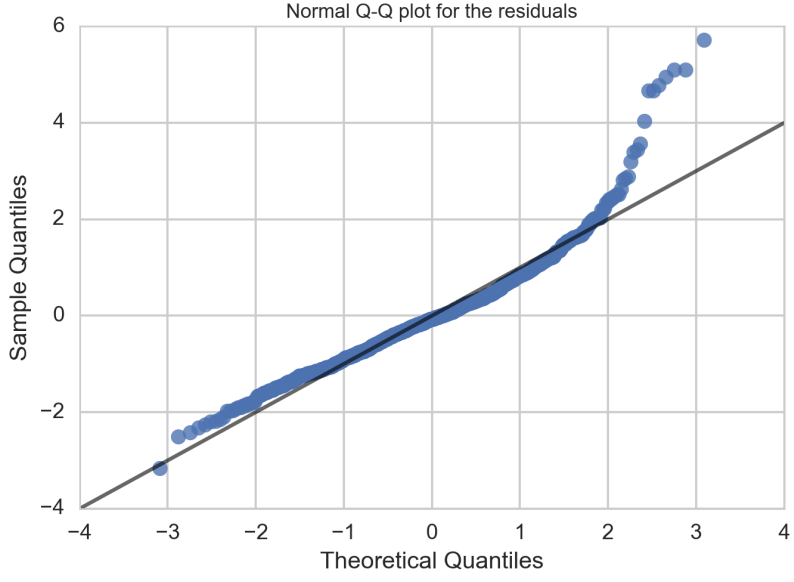
Assumption checking: fitted values against squared residuals



Assumption checking: residual distribution







Example: Direct Marketing

The diagnostics therefore reveal three problems:

- The residuals follow a nonlinear pattern. Hence, Assumptions 1 and 2 cannot be correct.
- The residuals have non-constant variance. Assumption 3 is not correct.
- The residuals are positively skewed. Skewed errors do not violate any assumptions, but are not ideal.

Data transformation

Data transformation

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

1. Modelling nonlinearity.
2. Meeting the assumption of constant error variance.
3. Reducing skewness.
4. Interpretability.

Log transformation

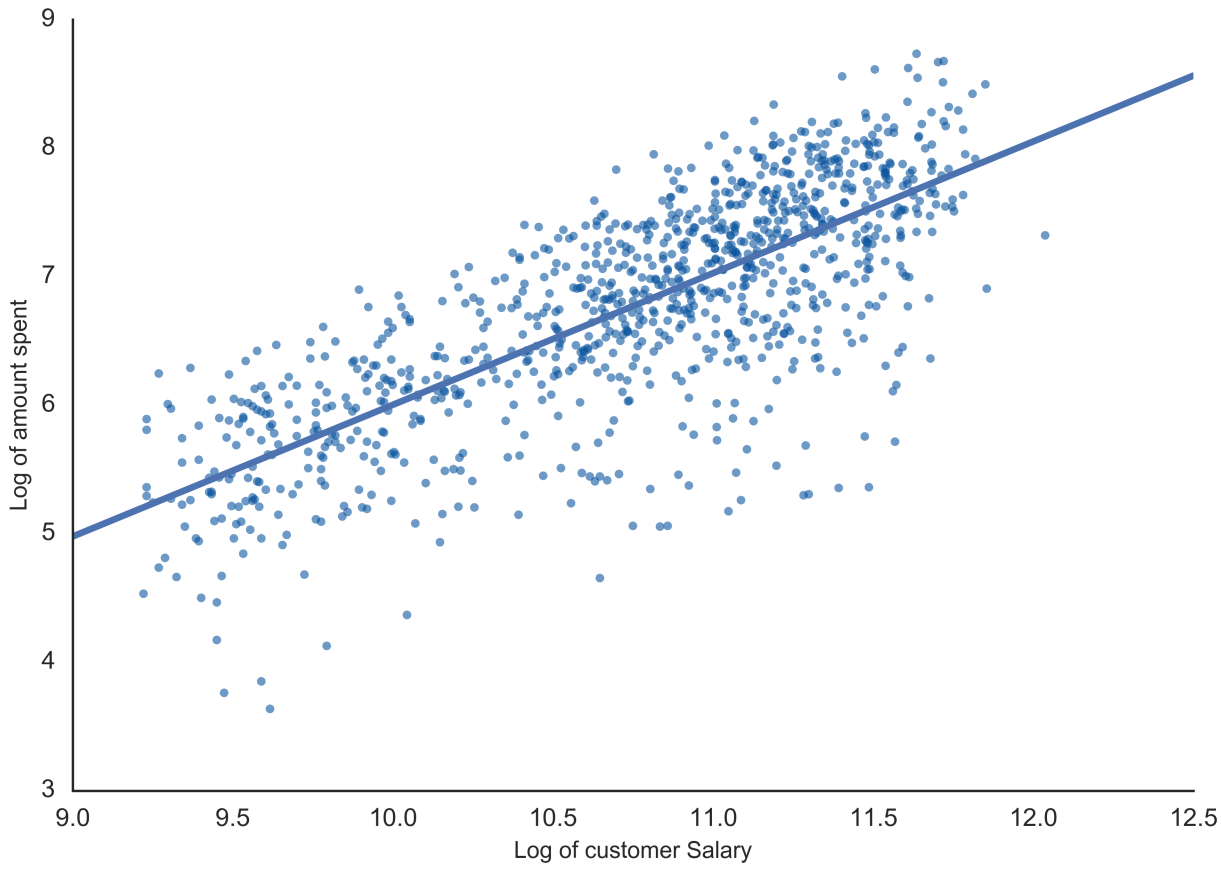
We can improve the model from the last section by considering the following specification:

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

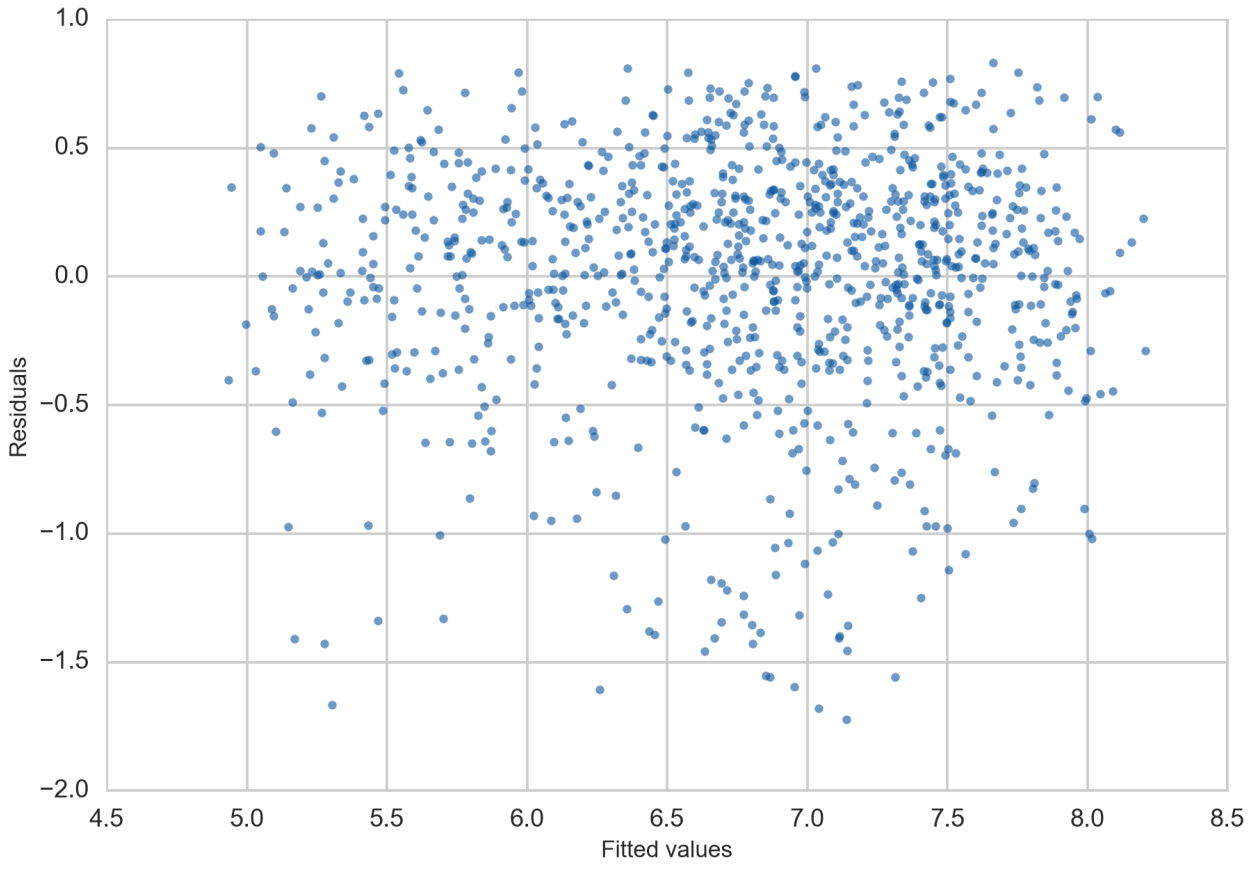
where $\log(\cdot)$ is the natural logarithm.

We now estimate the specification and look at the diagnostics.

Estimated regression (salary only)



Estimated regression: fitted values against residuals



Example: Direct Marketing

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

The diagnostics tell us that Assumptions 1–3 are much more reasonable in this case, so that we should prefer this specification.

Questions:

- 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 (key concept)

Models with a log transformation of the response are particularly relevant for business applications:

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

where $X_1, \ldots, X_p$ are any predictors.

This model means that for given predictor values,

$$\begin{aligned} Y &= \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon\right) \\ &= \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right) \exp(\varepsilon) \end{aligned}$$

Log transformation of the response

Assume that the errors are independent from X. Then,

$$E(Y|X = \boldsymbol{x}) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right) E(\exp(\varepsilon))$$

Interpretation of coefficients

Suppose that we compare the expectation with another case from the population where $X_1 = \boldsymbol{x}_1 + 1$. Then

$$E(Y|X = \boldsymbol{x} + \boldsymbol{e}_1) \exp(\beta_1) \times E(Y|X = \boldsymbol{x}),$$

where $\boldsymbol{e}_j$ indicates a unit vector (this makes the notation compact).

If β_1 is not far from zero,

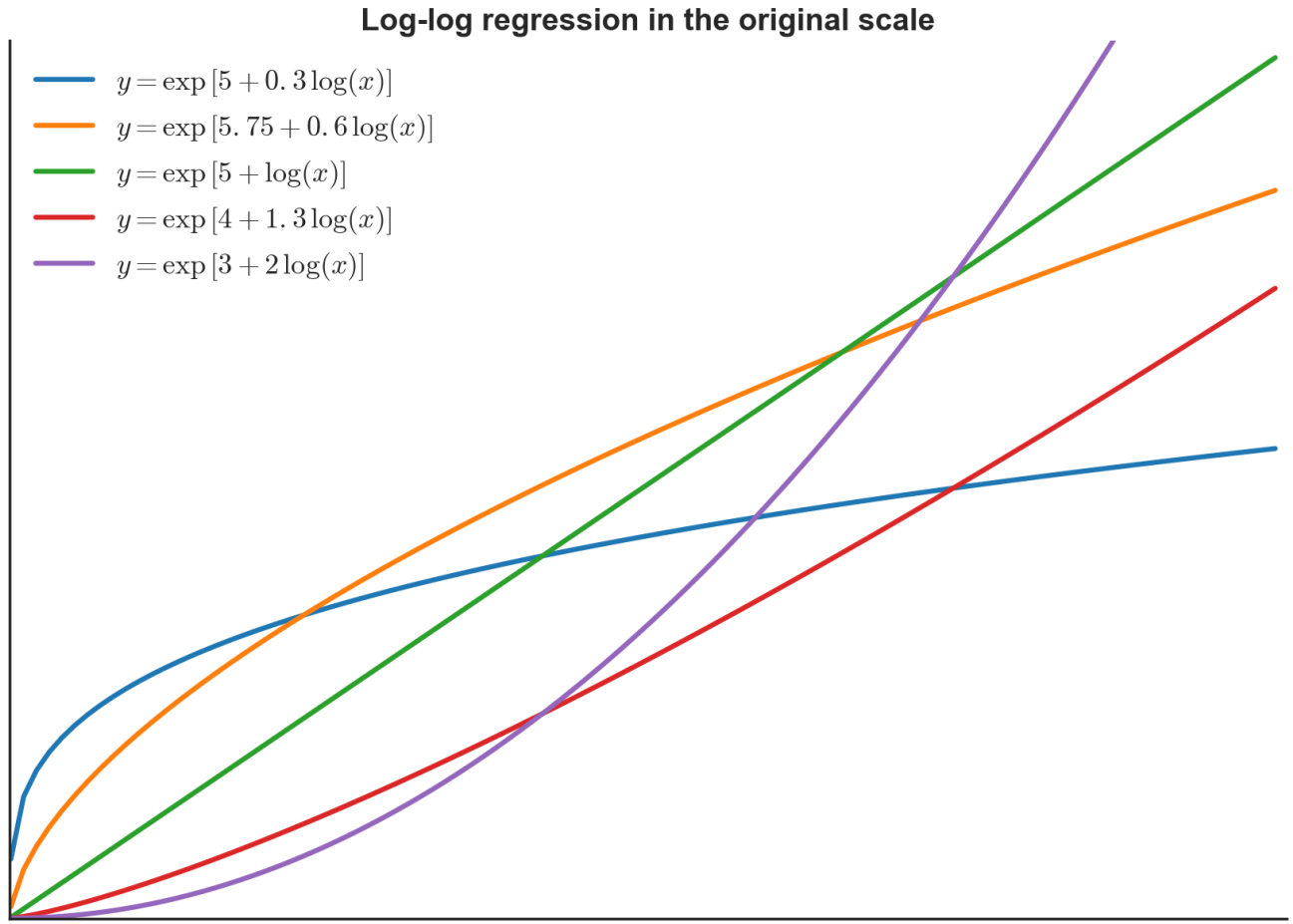
$$\exp(\beta_1) \approx 1 + \beta_1$$

That is, the expected value is approximately $100 \times \beta_1\%$ larger for $X_1 = \boldsymbol{x}_1 + 1$.

Interpretation of coefficients log transformations (key concept)

Case	Regression Specification	Interpretation
Linear-Log	$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$	A 1% difference in X is associated with a $0.01\beta_1$ expected difference in Y.
Log-Linear	$\log(Y) = \beta_0 + \beta_1 X + \varepsilon$	A unit difference in X is associated with a $100 \times \beta_1\%$ expected difference in Y.
Log-Log	$\log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon$	A 1% difference in X is associated with a $\beta_1\%$ expected difference in Y. β_1 is the elasticity.

Log transformations lead to nonlinear models



Estimating the conditional expectation

$$E(Y|X = \boldsymbol{x}) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right) E(\exp(\varepsilon))$$

A result known as **Jensen's inequality** tells us that

$$E(\exp(\varepsilon)) > \exp(E(\varepsilon)) = 1.$$

Therefore, the back transformation $\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)$ is downward biased for $E(Y|X = \boldsymbol{x})$.

Estimating the conditional expectation (key concept)

Duan (1983) proposed the following estimator:

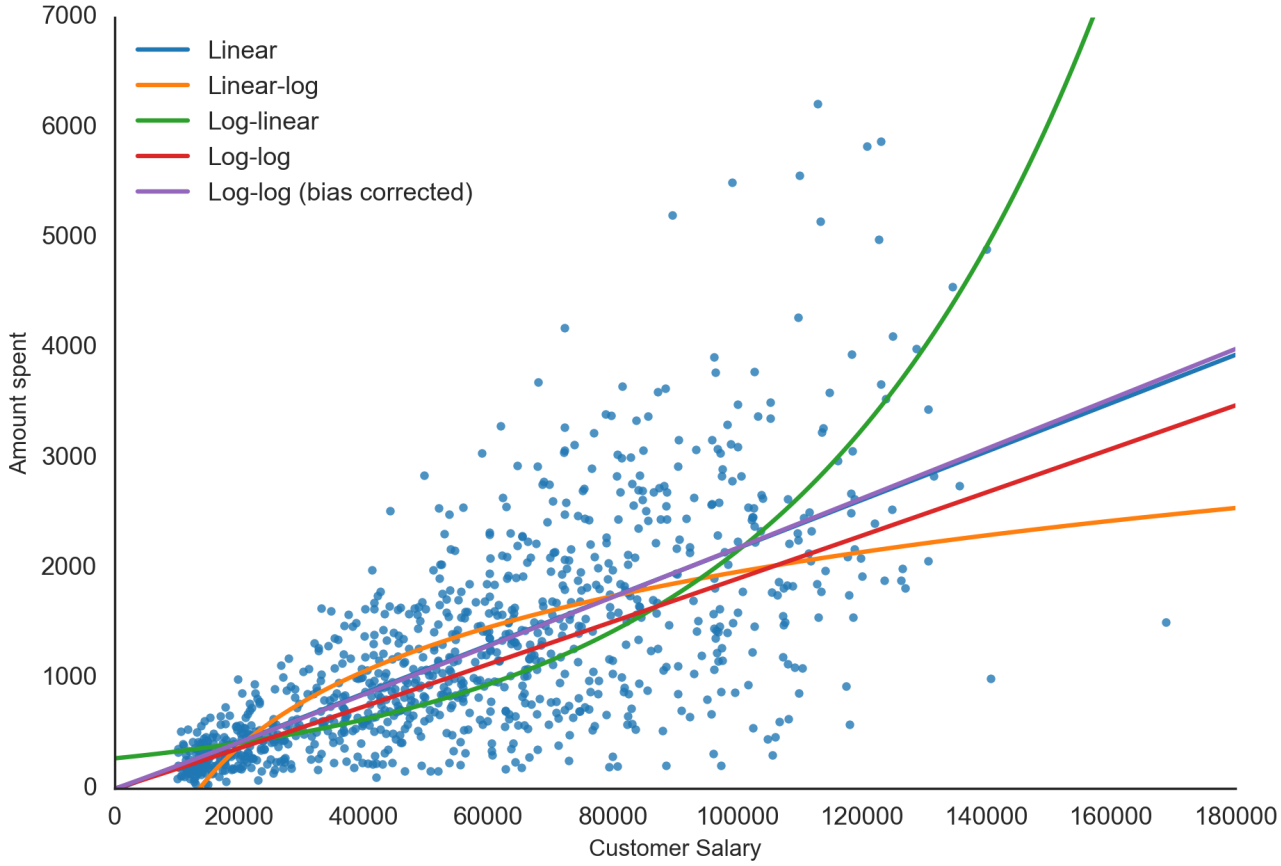
$$\widehat{y}(\boldsymbol{x}) = \exp\left(\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_j\right) \left[(1/N)\sum_{i=1}^{n} \exp(e_i)\right],$$

where e_i is the residual for observation i.

If the errors are normal then $E(\exp(\varepsilon) = \exp(\sigma^2/2)$, so that in this case we can use

$$\widehat{y}(\boldsymbol{x}) = \exp\left(\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_j + \widehat{\sigma}^2/2\right)$$

Example: Direct Marketing



Notice how bias correcting the re-transformation makes a large difference in the log-log estimate.

Power transformations

Other common types of transformations are power transformations.

Square root transformation. $\sqrt{y}$.

Inverse transformation. $1/y$.

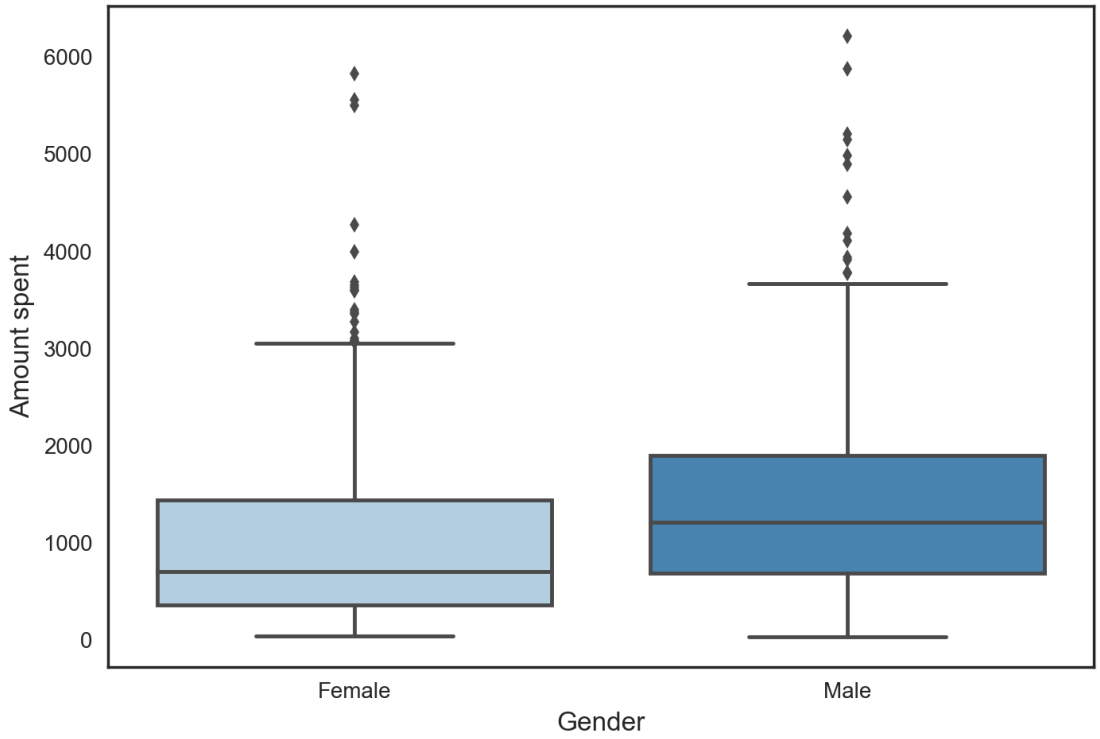
Square transformation. y^2.

Box-Cox. $(y^\lambda - 1)/(\lambda)$, for a parameter λ. The log transformation is a special case with $\lambda = 0$.

Categorical predictors

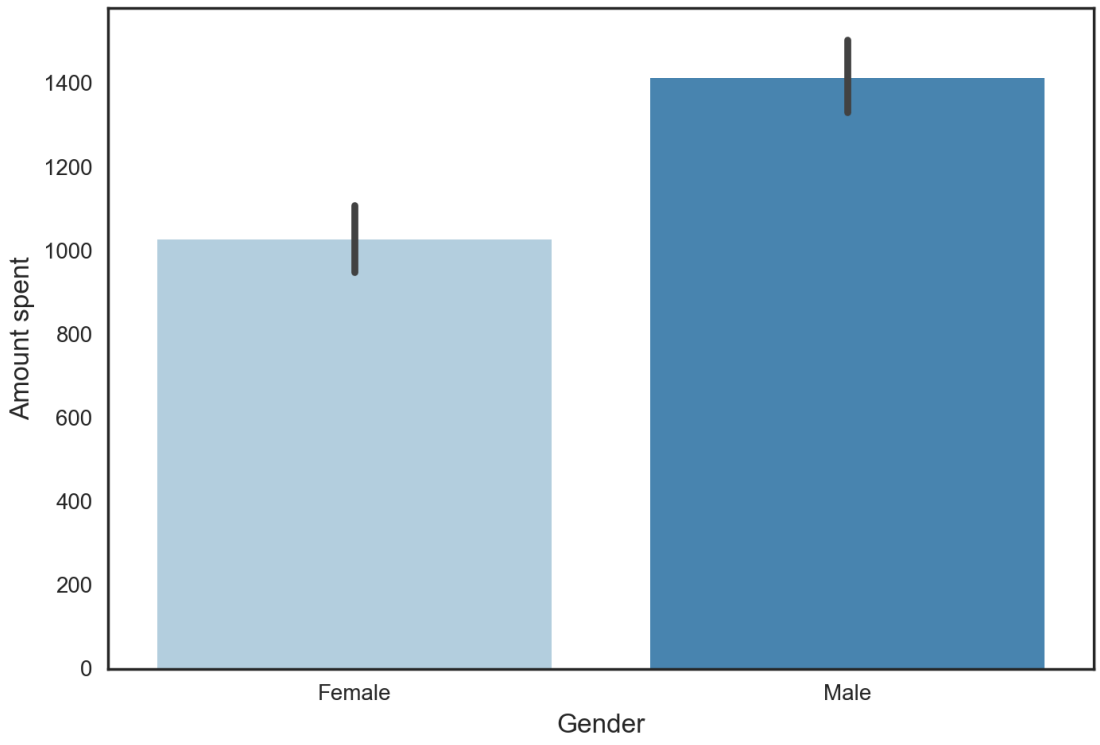
Categorical predictors

Several predictors in the dataset are **categorical variables**. We now study how to incorporate this type of variable into a regression model.

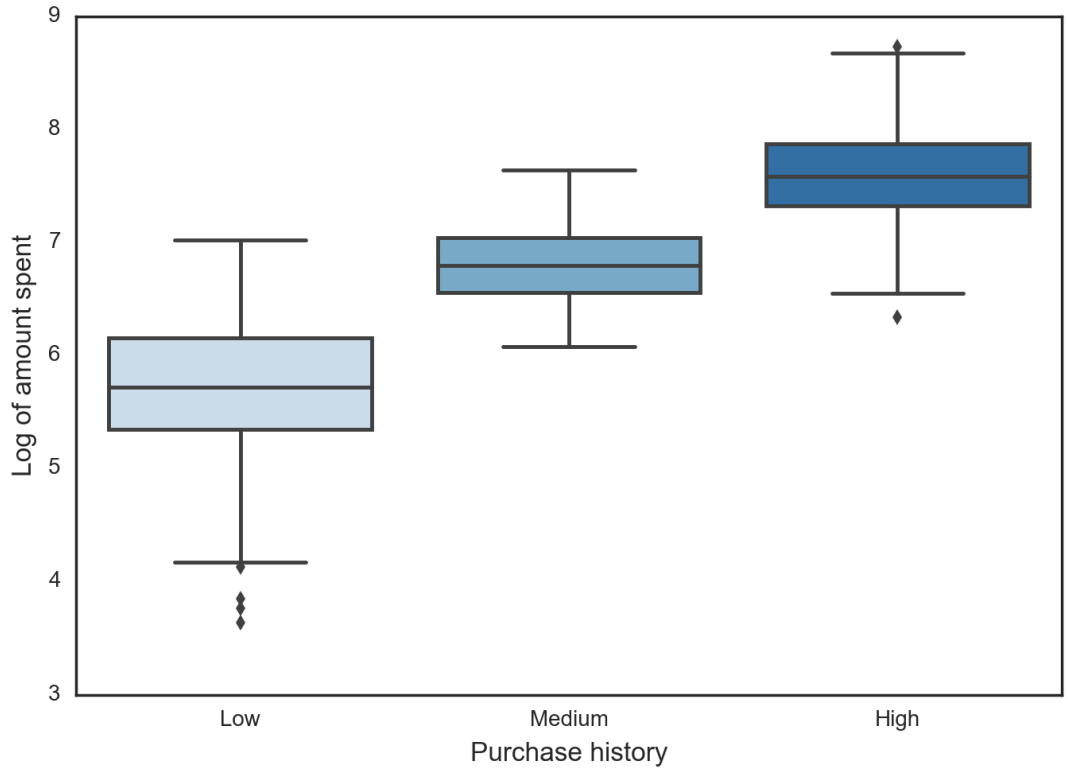


Exploratory plots

Another option is a bar plot:



Exploratory plots



Dummy variables (key concept)

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 one 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 one. Hence, we have to choose only one to use as predictor in the regression to avoid perfect multicollinearity.

Indicator variables

An **indicator variable** is a convenient notation for defining a dummy variable. Here are some examples:

$$I(\text{Single}) = \begin{cases} 1 & \text{if single,} \\ 0 & \text{otherwise.} \end{cases}$$

$$I(X > 0) = \begin{cases} 1 & \text{if } X > 0, \\ 0 & \text{if } X \leq 0. \end{cases}$$

Coefficient interpretation (key concept)

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.

Interpretation:

The coefficient β_1 the expected difference in Y when we compare two units which have the same value for all variables except that they belong to different categories.

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

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

Omnibus:	149.496	Durbin-Watson:	2.021
Prob(Omnibus):	0.000	Jarque-Bera (JB):	230.775
Skew:	-1.009	Prob(JB):	7.72e-51
Kurtosis:	4.210	Cond. No.	336.

Example: Direct marketing

$$\widehat{\log(\text{AS})} = \underset{}{-} \underset{(0.249)}{3.76} + \underset{(0.024)}{0.917} \times \log(\text{Salary}) + \underset{(0.002)}{0.048} \times \text{Catalogs} - \underset{(0.031)}{0.058} \times \text{Female}$$

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 5.8% less.

However, the coefficient is not statistically significant ($\alpha = 0.05$), so that we cannot reliably say that male and female have different average spending patterns conditional on salary and catalogs.

Binary categories

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

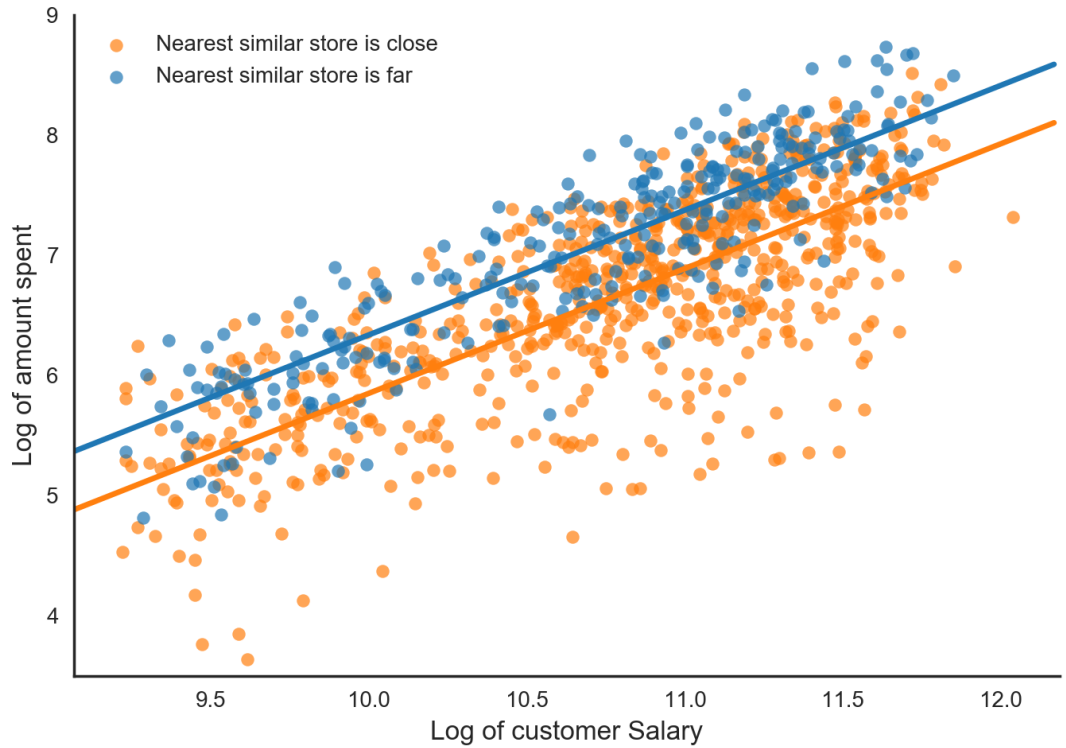
Another way to interpret the model is to think of it as specifying different intercepts depending on the class.

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

Example: Direct Marketing

Consider the model:

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



The lines are parallel.

Multiple categories (key concept)

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

Multiple categories

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

We then need to create 2 dummy variables. For example, we choose can *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

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_{m-1} D_{m-1} + \beta_m X_m + \ldots + \beta_p X_p + \varepsilon,$$

Interpretation:

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

$$\begin{aligned} \beta_1 = & E(Y|C = 1, X_m = x_m, \ldots, X_p = x_p) \\ & - E(Y|C = m, X_m = x_m, \ldots, X_p = x_p) \end{aligned}$$

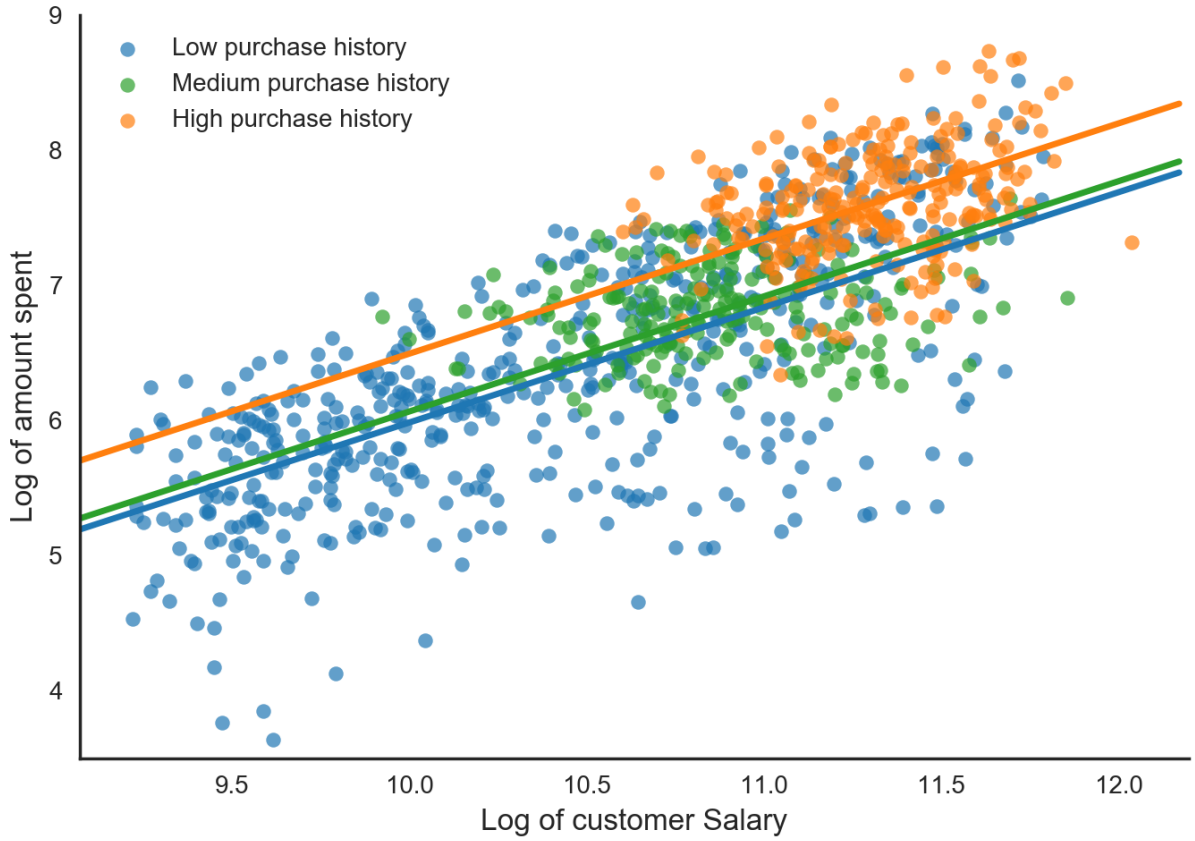
Multiple categories

$$\widehat{\log(\text{AS})} = \underset{(0.282)}{-2.91} + \underset{(0.0247}{0.834} \times \log(\text{Salary}) + \underset{(0.002)}{0.044} \times \text{Catalogs}$$
$$\underset{(0.039)}{- 0.01} \times \text{Medium} + \underset{(0.034)}{0.314} \times \text{High}$$

Interpretation:

If we select two customers with the same salary and number of catalogs sent, but one has a high purchase history and the other low, we expect the customer with high purchase history to spend 31.4% more.

Example: Direct Marketing



Interactions

Interaction modelling (key concept)

Suppose that we have a categorical predictor C and a continuous predictor X. In interaction modelling, we allow the relationship between Y and X to change depending on the C class.

For example, with $m = 2$ two classes and one quantitative predictor:

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

With $m = 3$ two classes:

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

Interactions

In a general setting in which C has m classes and interacts with a quantitative predictor X_j, we include $(D_1 \times X_j)$, $(D_2 \times X_j)$, $(D_{m-1} \times X_j)$ as predictor variables in the model.

Example: Direct Marketing

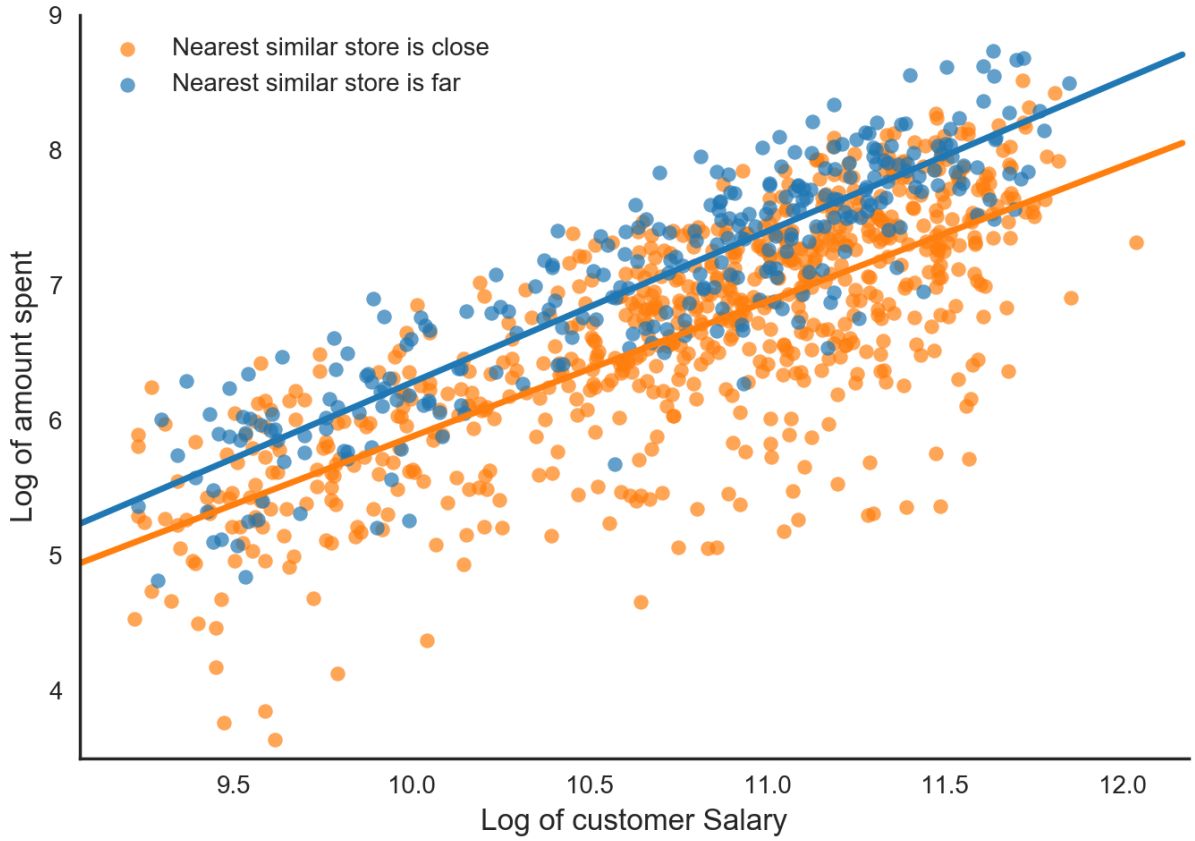
$$\widehat{\log(\text{AS})} = \underset{(0.279)}{-3.06} + \underset{(0.026}{0.848} \times \log(\text{Salary}) + \underset{(0.002)}{0.044} \times \text{Catalogs}$$
$$+ \underset{(1.27)}{2.86} \times \text{High} - \underset{(0.112)}{0.226} \times \text{High} \times \log(\text{Salary})$$

This as a regression with different intercept and salary slopes conditional on the purchase history category:

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

Interactions

Interactions



Interactions

- Interactions greatly increase the number of potential regressors in a model. It is not good practice to add interactions to a model indiscriminately.

- Rather, knowledge of the business problem, context, and exploratory data analysis may suggest the inclusion of interactions.

Polynomial regression

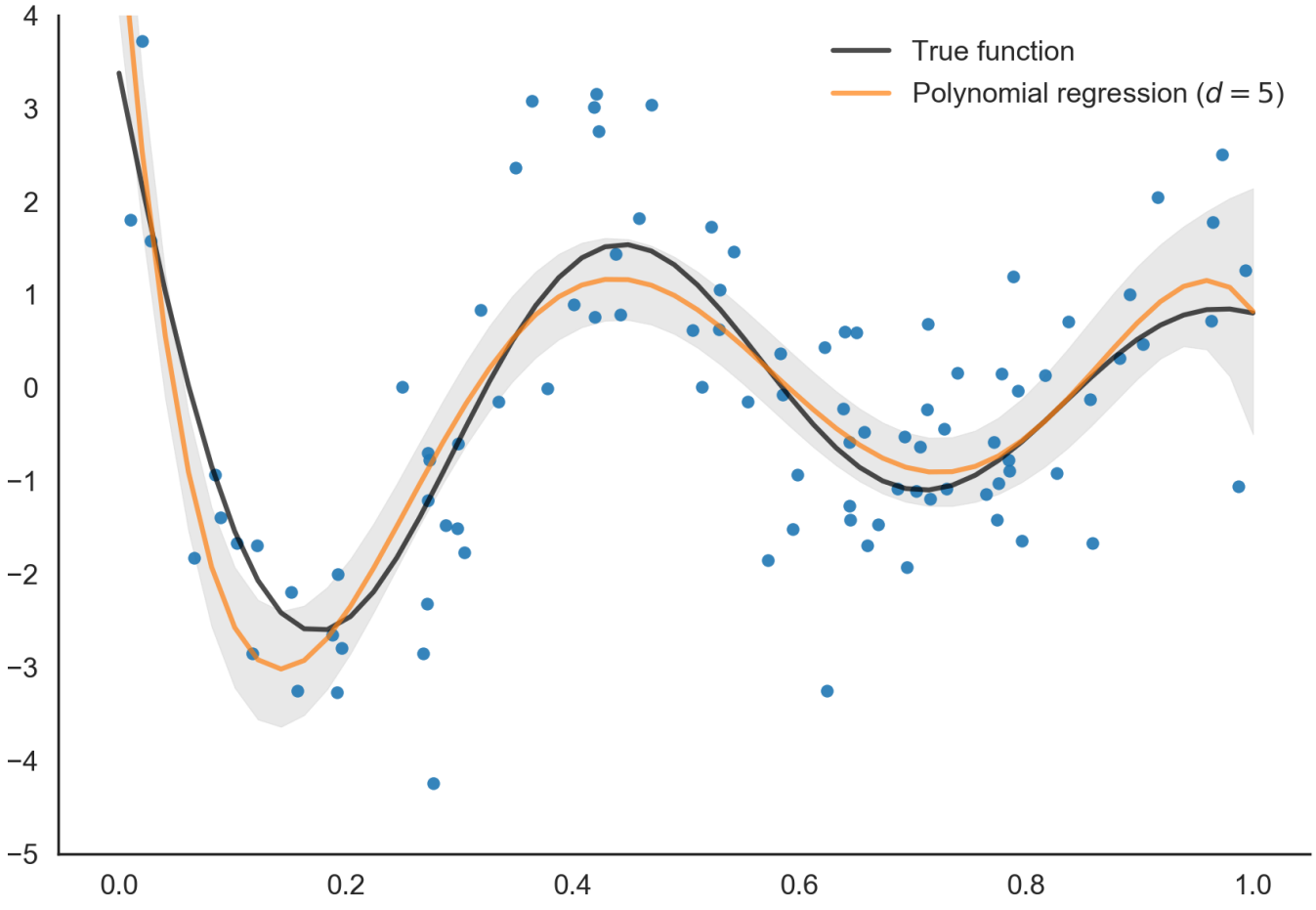
Polynomial regression (key concept)

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 polynomial degree. This is a MLR model and the 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.

Limitations

- Polynomials can overfit. A polynomial of degree d can approximate d points exactly, so that 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.

- You should extrapolate polynomial regressions to generate predictions outside the observed range of the predictor.

Bivariate polynomials

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 a bivariate polynomial regression as approximation to $f(X_1, X_2)$:

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

Bivariate polynomials and interaction terms

Bivariate polynomial regression:

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

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, but may be less justified.

Review questions

- Why is EDA important for supervised learning?
- 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 specification?
- What is interaction modelling?
- What is a polynomial regression and why do we use it?