Regression and Logistic Regression and K-Nearest Neighbor Prediction

- Regression is a parametric method:
  - Parametric methods are rooted in ***specific assumptions***.
  - Their modeling outcomes can be generalized to an unknown population as long as their assumptions are satisfied. Thus the validity of the underlying assumptions need to be verified.
  - Implicitly the scale of the feature is accounted.
  - Aside from making predictions, parametric methods also allow to make statements about the underlying data generating process.
  - Due to the small number of parameters, parametric methods are rather inflexible.
- Non-parametric methods:
  - They are more ***data driven*** than relying on assumptions.
  - They are more flexible to adjust to an underlying pattern in the sample data.
  - The sole objective of non-parametric methods in ML is prediction.
  - The scale of the features needed to be handled explicitly.

Parametric Linear Regression

- The parameters in multiple linear regression model $y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_K \cdot x_{iK} + \varepsilon_i$ are the regression coefficients $\beta_0, \beta_1, \dots, \beta_K$.
- The predicted value is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i1} + \cdots + \hat{\beta}_K \cdot x_{iK}$ where $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K$ are the estimated regression coefficients.
- These parameters are estimated by a method called **ordinary least squares**, which aims at finding that set of the parameters $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K$ which ***minimize*** the residual sum of squares RSS, i.e.,

$$\min_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_K} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

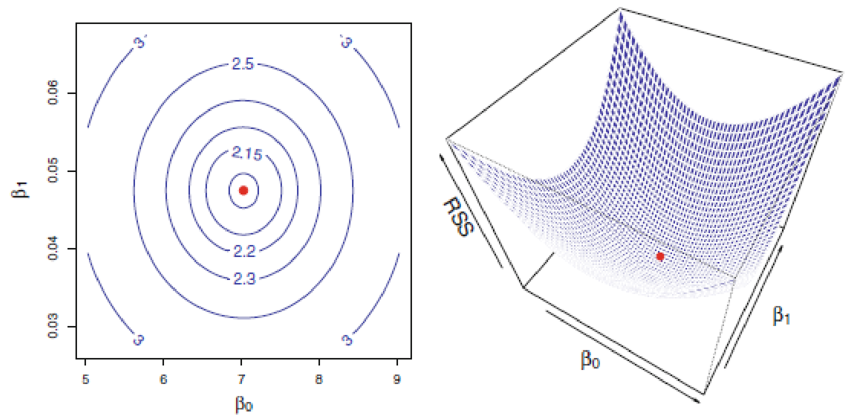
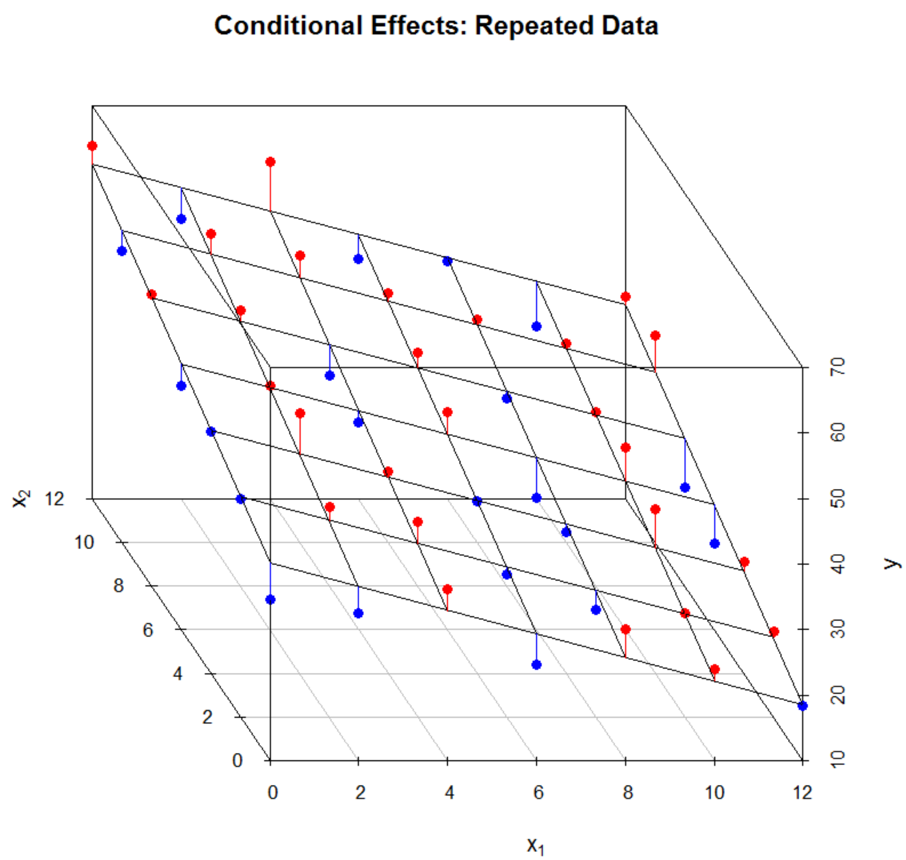


FIGURE 3.2. *Contour and three-dimensional plots of the RSS on the* `Advertising` *data, using* `sales` *as the response and* `TV` *as the predictor. The red dots correspond to the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, given by (3.4).*

- For two independent variables X_1 and X_2 the model has the graphical representation:



- The estimate parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$ internally account for the scale of the features $X_1, \ldots, X_K$.
- Assumptions about the model structure:
  - [A1] The features X are free of random effects.
  - [A2] The error term has an expected value of zero, i.e., $E[\varepsilon_i] = 0$
  - [A3] All relevant features are in the model.

  - [A4] The underlying data generating process is linear in the features.
  - [A4] The variance of the error term is constant, i.e., $Var[\varepsilon_i] = constant\ \forall\ i$
  - [A5] The error terms are independent among each other, i.e., $Cov[\varepsilon_i, \varepsilon_j] = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$.
  - [A6] The error term is normally distributed $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)\ \forall\ i$.
- When these assumptions are satisfied, the estimated regression parameter $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$ are unbiased with the smallest standard errors. Thus, the estimate model can be generalized to yet not available data points.

Addressing questions about the model

1. Is at least one feature relevant in predicting the target? (→ global F-test)
2. Do all features or just a selected set help explaining the target? (→ t-test and stepwise regression)
3. How well does the model fit the data? (→ R^2_{adj} or AIC)
4. How do we handle uncertainty in the prediction? (→ prediction confidence intervals)

- The **global $\boldsymbol{F}$-test** allows to evaluate whether the model overall has some explanatory power, i.e.,

$$H_0: \beta_1 = \cdots = \beta_K = 0 \ \text{ against at least one } \ \beta_k \neq 0$$

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

- Each feature can be tested whether it is relevant in explaining a proportion of the variation in the target by the statistical test by the ***t*-test**:

$$H_0: \beta_k = 0 \ \text{ against } \ H_1: \beta_k \neq 0$$

If the associate error probability of rejection the null hypothesis H_0 – even though it is true – becomes reasonable small we accept the alternative hypothesis H_1.

	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

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

- **Forward stepwise selection** of a set of relevant features allows to heuristically identify a set of relevant features:

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

- Alternatively, backward selection procedures or mixed procedures can be employed.
- The **overall explanatory power** of the model in terms of explained variation of the target variable is measured by the adjusted R^2_{adj}, i.e.:

$$R^2_{adj} = 1 - \frac{RSS/(n-K-1)}{TSS/(n-1)}$$

It penalizes for the complexity of the model (increase in the variance for the MSE).

- An alternative goodness of fit measure is the Akaike Information Criterion. It becomes for normal distributed error terms:

$$AIC = \frac{1}{n \cdot \hat{\sigma}^2} \cdot (RSS + 2 \cdot K \cdot \hat{\sigma}^2)$$

A small AIC is preferred. It penalizes and overfitted model more than the R^2_{adj}.

- Regression is a statistical model involving an error distribution. The error distribution is associated with the irreducible error of the model. ***Confidence intervals*** around the regression plane or an individual point prediction allows to assess the predictive quality of the model.

Flexibility of the regression model

- Categorical features (see also Boehmke p 61)
  - Beside metric features regression can also handle factors (categorical features).
  - Each factor level is encoded as a dummy variable.
  - Due to the redundancy of the set of factor levels one factor level needs to be dropped explicitly from the model. It can be calculated implicitly.
- Non-linear functions in the features
  - Allows expressing non-linear relationships between the target and the features in a linear setting.
  - Each feature can be transformed, e.g., Box-Cox or Yeo-Johnson.
  - Each feature can be expressed as a polynomial function, i.e.,

$$\beta_{k_1} \cdot X_k + \beta_{k_2} \cdot X_k^2 + \beta_{k_3} \cdot X_k^3 + \cdots$$

  - Polynomial functions bear the risk of overfitting the data.

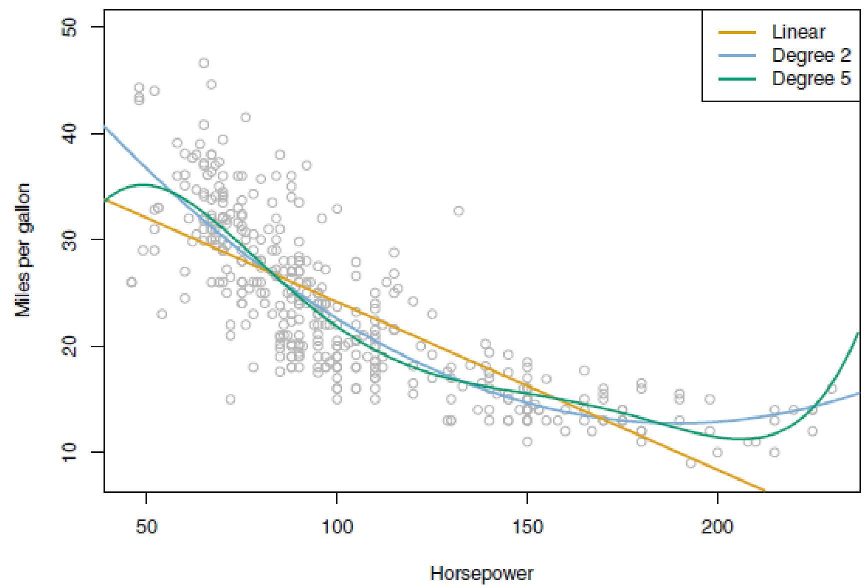


FIGURE 3.8. *The* Auto *data set. For a number of cars,* mpg *and* horsepower *are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes* horsepower^2 *is shown as a blue curve. The linear regression fit for a model that includes all polynomials of* horsepower *up to fifth-degree is shown in green.*

- Interaction effects
  - Features many influence a target in unison rather than separately. One feature may enhance or diminish the effects of another variable.
  - This interplay among features is modelled by interaction effects, e.g.,

$$\begin{aligned} Y &= \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1 \cdot X_2 + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \cdot X_2) \cdot X_1 + \beta_2 \cdot X_2 + \varepsilon \end{aligned}$$

	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

TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

Caveats of Regression

- As soon as the target variable is non-linearly transformed, the regression model becomes non-linear. It still can be evaluated by conditional effects plots.
- Outlying observations must be identified and handled with care because they exhibit a strong influence on the estimated parameters $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K$.
- Highly correlated features are redundant. This redundancy increases the uncertainty (standard error) in the estimated parameters $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K$.
- Autocorrelation and heteroscedasticity leave the estimated parameters $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K$ unbiased but usually inflate their standard errors.

Parametric Logistic Regression

- Logistic regression is a parametric **supervised** classification procedure.
- The target variable is a factor (categorical variable) describing the mutually exclusive and exhaustive class membership of each observation.
- The objective is to predict the class membership probabilities for each observation. Overall possible classes these probabilities sum to one. demo
- For a binary (just two categories) target variable the target variable becomes

$$Y_i = \begin{cases} 1 & event\ happening \\ 0 & event\ not\ happing \end{cases}$$

and the predicted value given at a given set features becomes

$$\hat{p}_i = \Pr(Y_i = 1 | x_{i1}, \dots, x_{iK}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i1} + \dots + \hat{\beta}_K \cdot x_{iK})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i1} + \dots + \hat{\beta}_K \cdot x_{iK})}$$

- Per standard assumption $\Pr(Y_i = 1 | x_{i1}, \dots, x_{iK})$ follows a binomial distribution with an associated likelihood function
- Numerical optimization finds the estimated parameters $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K$.
- The probabilities are inherently non-linear with respect to $X_1, \dots, X_k$

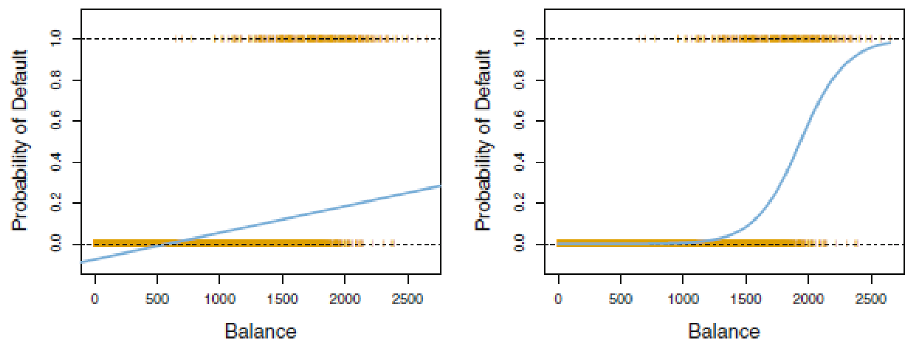


FIGURE 4.2. *Classification using the* Default *data.* Left: *Estimated probability of* default *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* default*(*No *or* Yes*).* Right: *Predicted probabilities of* default *using logistic regression. All probabilities lie between 0 and 1.*

but after the transformation $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i1} + \dots + \hat{\beta}_K \cdot x_{iK}$ the logits $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$ becomes are linear function in $X_1, \dots, X_k$.

- The estimated parameters $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K$ again capture the varying scales of $X_1, \dots, X_k$.
- Feature that are based on factors, interaction effects and polynomial specifications can be easily accommodated.

Non-parametric k-nearest Neighbors

data-driven, no estimated parameters

- k-nearest neighbors estimation of a metric or class feature is a non-parametric methods.
- It is only driven by the hyper-parameter k which cannot be estimated from the data.
- In order to calculated among objects distances, the scale of metric features needs to be set by the analyst perhaps by making the feature scales comparable.
- The definition of object distances in terms of categorical features is ambiguous.
- Irrespectively of whether the target Y is metric or categorical, the underlying predicted value $\hat{Y}_0$ at location $X_{01}, \dots, X_{0K}$ is

$$\hat{Y}_0 = \frac{1}{k} \cdot \sum_{X_{i1},\dots,X_{iK} \in \mathcal{N}_0} Y_i$$

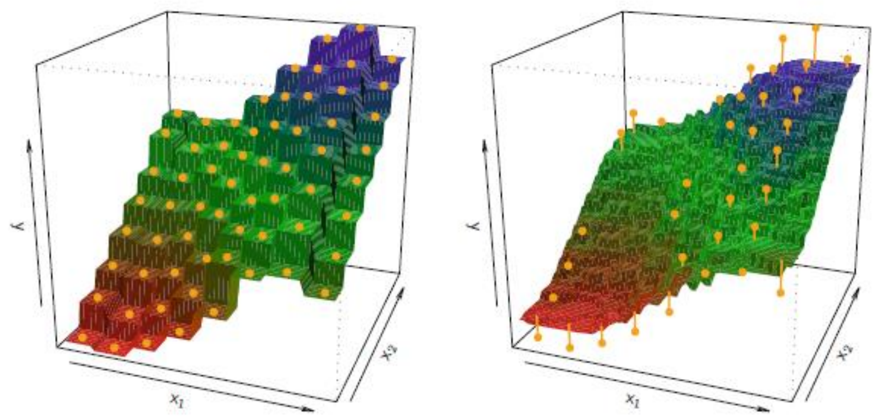


FIGURE 3.16. *Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots).* Left: *$K = 1$ results in a rough step function fit.* Right: *$K = 9$ produces a much smoother fit.*

- For $k = 1$ the KNN fits the sample observations perfectly (most flexible fit). The bias is low but the sample-to-sample variance is high.