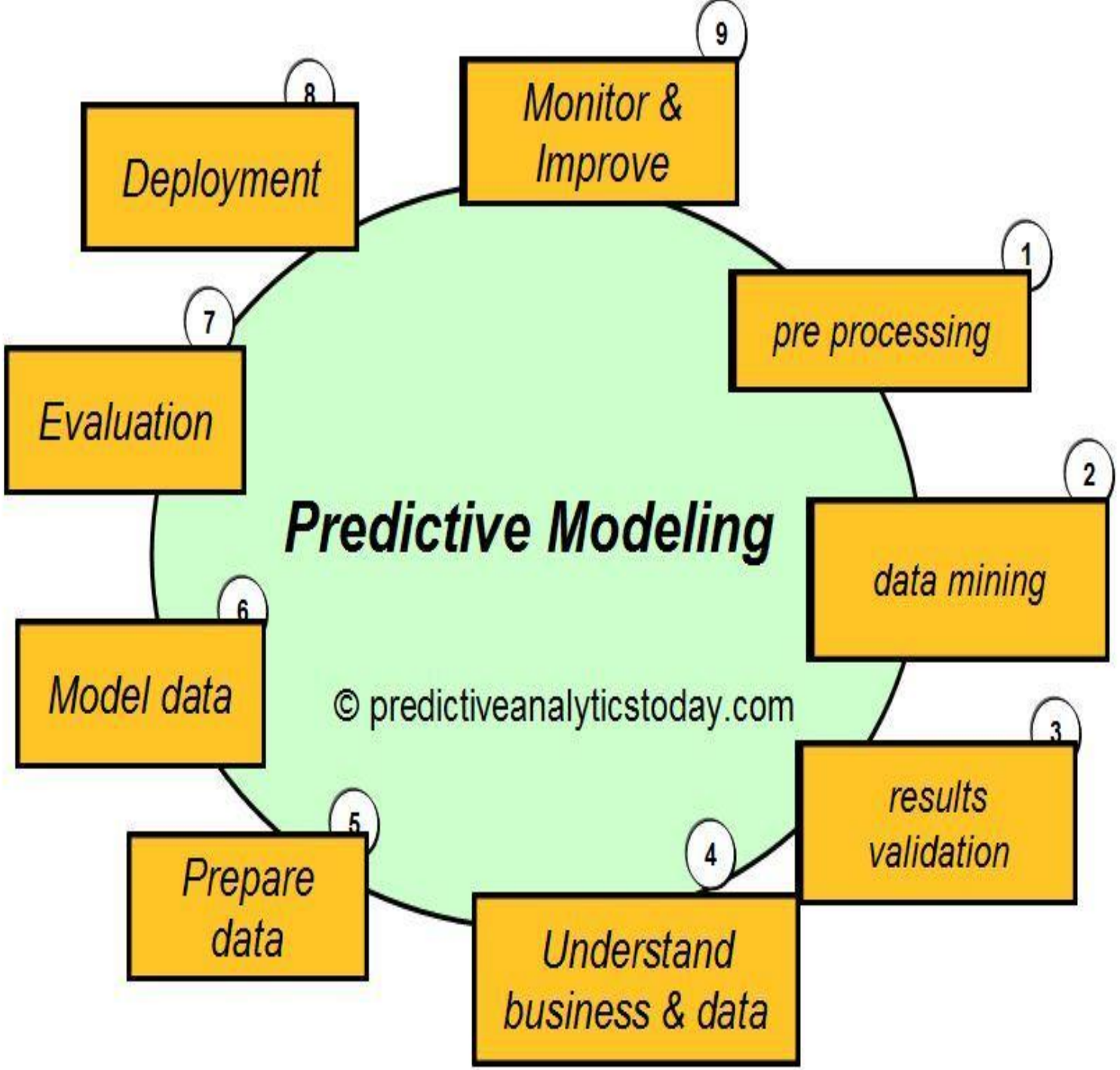


Module 1

INTRODUCTION TO PREDICTIVE MODELING

Introduction

- **Statistical learning** involves a wide array of techniques for interpreting data. These techniques are generally categorized into **supervised** or **unsupervised learning**.
- **Supervised learning** focuses on constructing a statistical model that predicts or estimates an output using one or more inputs.
- **Unsupervised learning** on the other hand only has input data but no supervising output. It is more for descriptive purposes, learning relationships and structures within the data.
- **Predictive modeling** is a form of supervised learning in which future outcomes are forecasted or predicted based on historical data.
- The goal of predictive modeling is to identify a function that effectively combines various measured explanatory variables to predict the outcome of a response variable accurately

Examples of predictive models

- Examining the influence of age, education, and calendar year on the wages of men in the Midwest, aiming to predict an individual's wage based on these factors.
- Using features like square footage, location, number of bedrooms and bathrooms, lot size, recent renovations, energy efficiency ratings to predict the sale price of homes.
- To tackle employee turnover challenges, the HR department of a mid-sized tech firm is implementing a predictive model to forecast the likelihood of employee attrition using various attributes; age, department, tenure (length of stay), job role, salary, education level, performance rating,work hours, job satisfaction

Importance of Predictive Modeling

Predictive modeling is invaluable in decision-making processes, as it allows organizations and individuals to anticipate future events and plan accordingly. For example,

- In business, predictive modeling can optimize marketing strategies, improve customer relationship management, and enhance operational efficiency.
- In healthcare, it aids in disease prediction and patient care management.
- In finance, it is used for risk assessment, e.g. fraud detection, credit scoring
- In meteorology for weather forecasting.

Its widespread application across different fields makes it significant in transforming data into actionable insights, leading to more informed and effective decisions.

Types of Predictive Models

1. Regression Models: Used for predicting an outcome.
    - Linear Regression (continuous and quantitative)
    - Logistic regression (binary or several classes)
2. Classification Models: Used for predicting categorical outcomes.
    - Decision Trees
    - Support Vector Machines (SVM)
    - Neural Networks
    - Linear discriminant analysis
    - Etc.
3. Generalized linear models
4. Time Series Models: Used for forecasting trends and patterns over time.
5. Clustering Models: Used for segmenting data into groups.

Terminologies

- In machine learning, data is typically divided into two sets: the training set, the test set / validation set.
  - A *training set* is a subset of the data that is used to develop the model; used to estimate the values needed in the model equation.
  - A *test set* is a subset of the data that is used to evaluate the performance of a final set of trained models. It is used for assessing how well the model generalizes to new, unseen data.
  - When dividing a dataset into training and test sets, common suggestions include allocating 60% for training and 40% for testing, 70% for training and 30% for testing, or 80% for training and 20% for testing.
- The terms *predictors, independent variables, attributes, or descriptors* refer to the data that serve as input in the predictive model.
- The *outcome, dependent variable, target, class,* or *response* denotes the variable that the prediction equation aims to forecast or predict.

Terminologies

- *Continuous* data refers to data that can take on any numeric value within a given range, representing measurements. Examples include height, weight, temperature, blood pressure and time.
- *Categorical data* is a type of data that can be divided into groups or categories that are typically qualitative in nature. value for the variable, and the categories are usually non-numeric. Examples include Blood Type (A, B, AB, O), gender, color, etc.

Predictive Modeling Process

1. Problem Definition: Clarify the objective and goals of the model, understanding the specific issue or question it aims to address.
2. Data Collection: Gather relevant data from various sources, ensuring it is sufficient and appropriate for the modeling task. Ensure that enough data is collected to train a robust model.
3. Data Preprocessing: This involves assembling the data into the appropriate form for the model.
   - Cleaning Data: Address missing values, remove duplicates, and correct inconsistencies.
   - Data Transformation: Normalize or standardize data to bring everything to a comparable scale.
   - Feature Engineering: Create new features from existing data to improve model performance, e.g he categorical data fields.
   - Data Reduction: Use techniques like dimensionality reduction to simplify the model without losing critical information.
   - Separate data into the desired training, testing, and validation sets.

Predictive Modeling Process

4. Model Selection: Choose an appropriate modeling technique (like regression, classification, clustering) based on the problem type and data characteristics.
5. Model Training:
    - Training Dataset: Use the prepared dataset to train the model. This involves feeding the data into the model so it can learn from it.
    - Parameter Tuning: Adjust the model parameters to optimize performance. This can be done through methods like grid search or random search.
    - Validation: Use a cross-validation to tune the hyperparameters and refine the model.
6. Model Evaluation: Assess the model's performance using a separate test dataset and relevant metrics to ensure accuracy and reliability.
7. Deployment: Implement the model in a real-world setting, integrating it into existing systems and monitoring its performance for ongoing effectiveness. Be prepared to retrain or update the model as new data becomes available or as the environment changes.

Model Assessment

1. **Regression Metrics**
    1. Mean Absolute Error (MAE)
    2. Mean Squared Error (MSE)
    3. R-squared
2. **Classification Metrics**
    1. Accuracy
    2. Precision, Recall, and F1 Score
    3. Confusion Matrix
    4. ROC Curve and AUC

Model Assessment

Regression Metrics

1. **Mean Absolute Error (MAE):** MAE is the average of the absolute differences between prediction and actual observation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i is the actual value of the response and $\hat{y}_i$ is the predicted value.

- An MAE of 0 means perfect predictions. As MAE increases, the model's predictions are less accurate.

2. **Mean Squared Error (MSE):** MSE is a measure of the average of the squares of the errors. It calculates the average squared difference between the estimated values and the actual value.

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

- MSE penalizes larger errors more severely than smaller ones, due to squaring each error. A MSE of 0 indicates perfect predictions. Higher values mean poorer model performance. MSE is more sensitive to outliers compared to MAE.

Model Assessment

Regression Metrics

3. R-squared (Coefficient of Determination)

R-squared is a statistical measure that represents the proportion of the variance for the dependent variable that's explained by the independent variables in a regression model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

where $\bar{y}_i$ is the mean of the observed data.

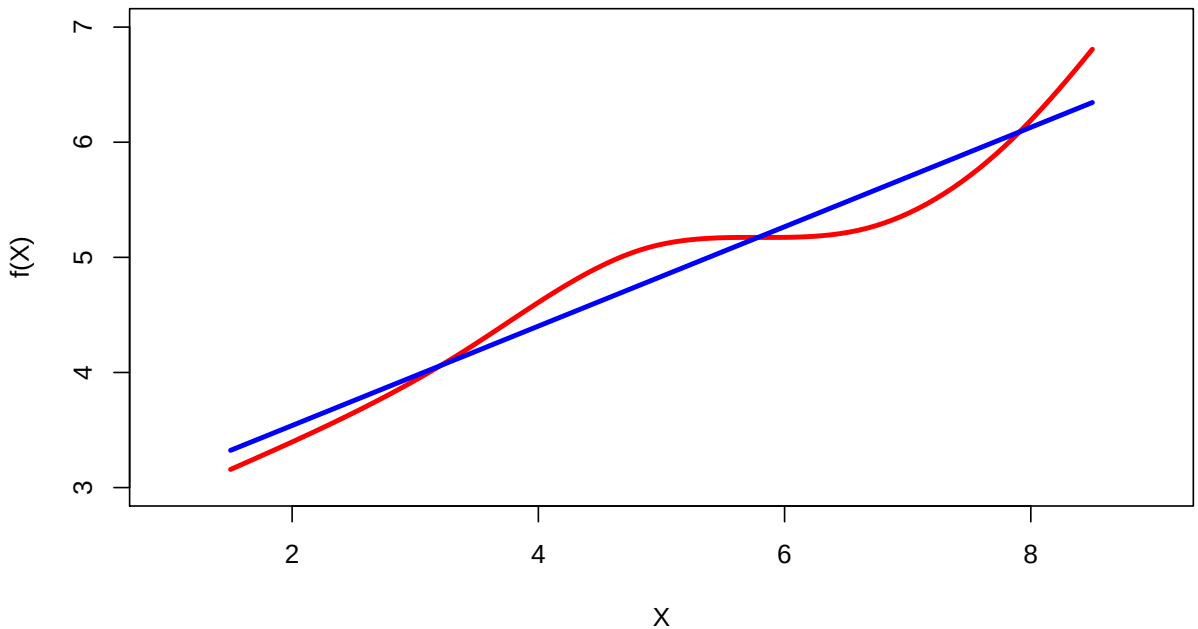
- R-squared values range from 0 to 1.
- A value of 0 indicates that the model explains none of the variability of the response data around its mean, while a value of 1 indicates that it explains all the variability.
- However, a high R-squared doesn't necessarily mean a good model fit (e.g., it can be artificially inflated with more variables).

UP NEXT!

Simple and Multiple 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.

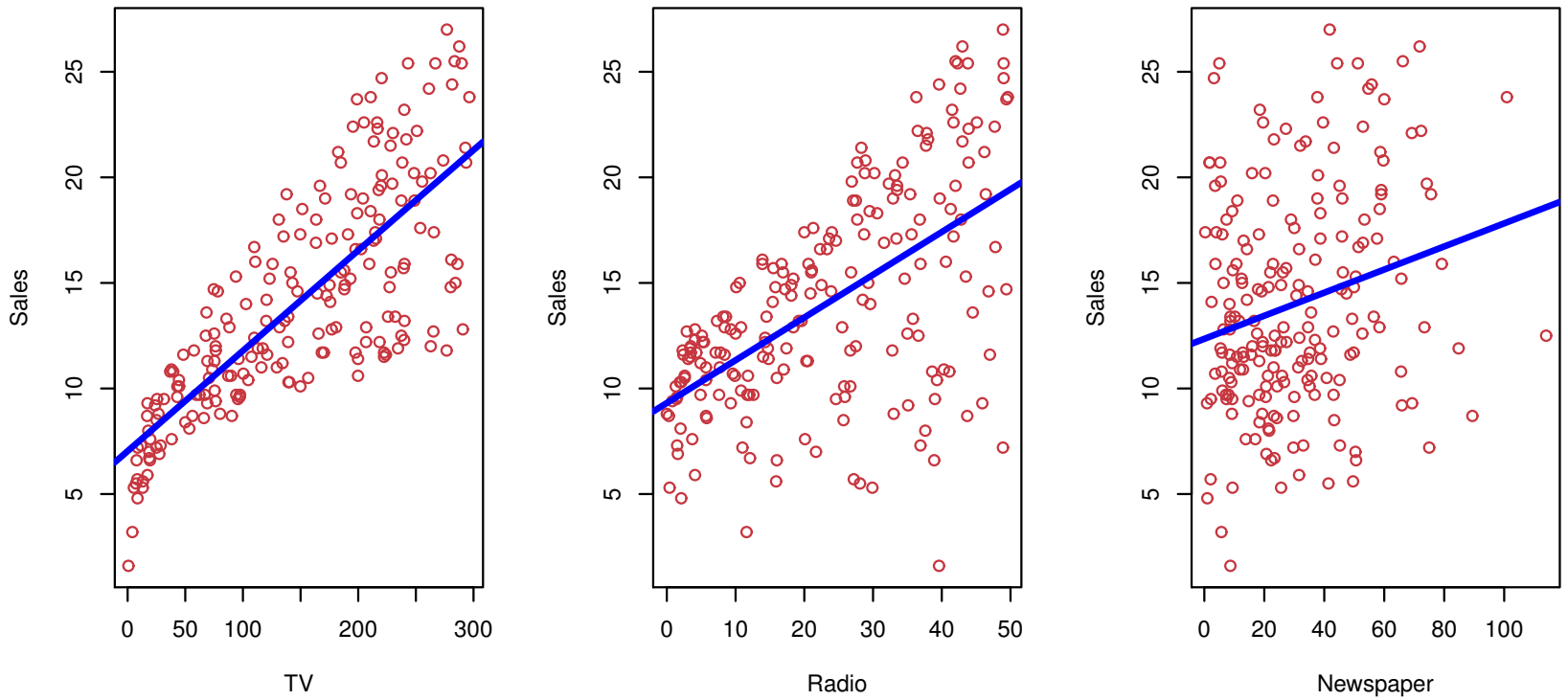
Linear regression for the advertising data

Consider the advertising data shown on the next slide.

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?

Advertising data



Simple linear regression using a single predictor X.

- 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 $\hat{\beta}_1$ for the model coefficients, we predict future sales using

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

  where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The *hat* symbol denotes an estimated value.

Estimation of the parameters by least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*
- We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

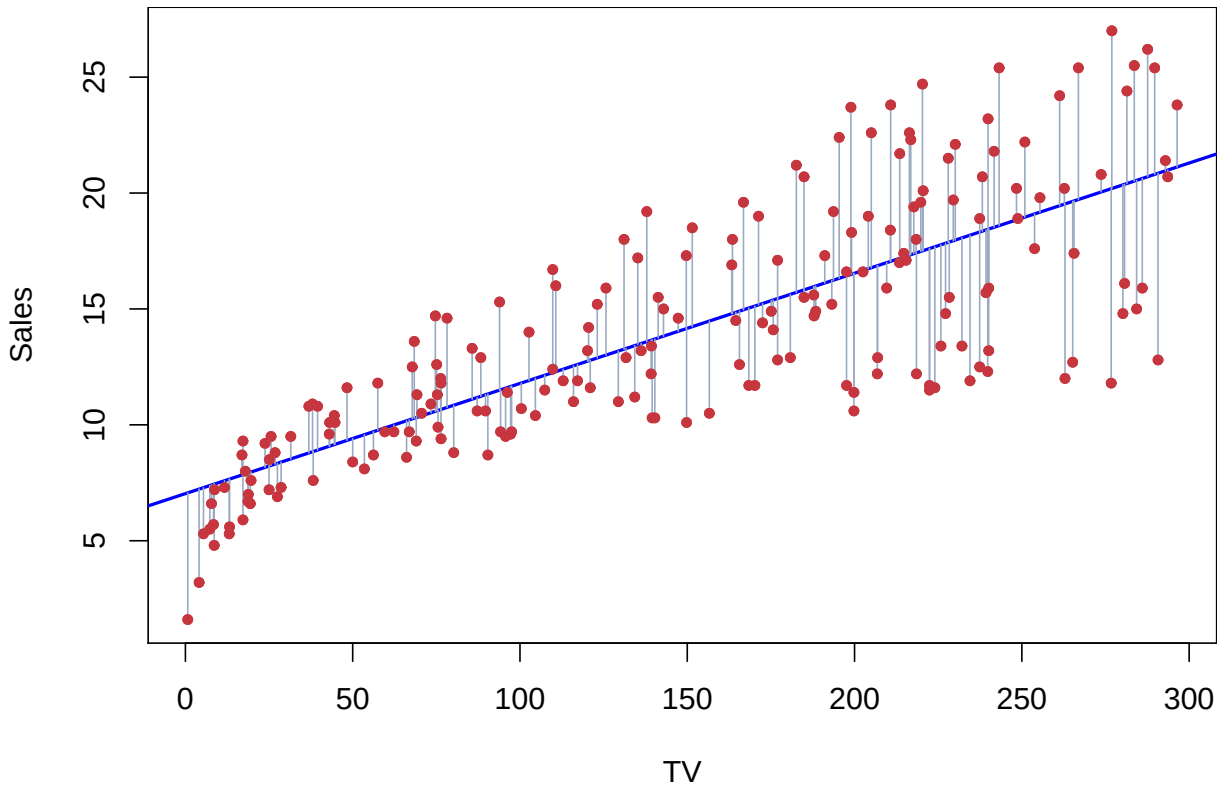
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

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

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

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

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. We have

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right],$$

  where $\sigma^2 = \mathrm{Var}(\epsilon)$

- 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

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

Confidence intervals — continued

That is, there is approximately a 95% chance that the interval

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

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$

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

  H_0 : There is no relationship between X and Y

  versus the *alternative hypothesis*

  H_A : There is some relationship between X and Y.

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

  versus

$$H_A : \beta_1 \neq 0,$$

  since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Hypothesis testing — continued

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

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

- 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*.

Results for the advertising data

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

Assessing the Overall Accuracy of the Model

- We compute the *Residual Standard Error*

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  where the *residual sum-of-squares* is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- *R-squared* or fraction of variance explained is

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

  where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.
- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y:

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

Advertising data results

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Multiple Linear Regression

- Here our model is

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

- 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

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated — a *balanced design*:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as *"a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed"*, are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous — when X_j changes, everything else changes.

Estimation and Prediction for Multiple Regression

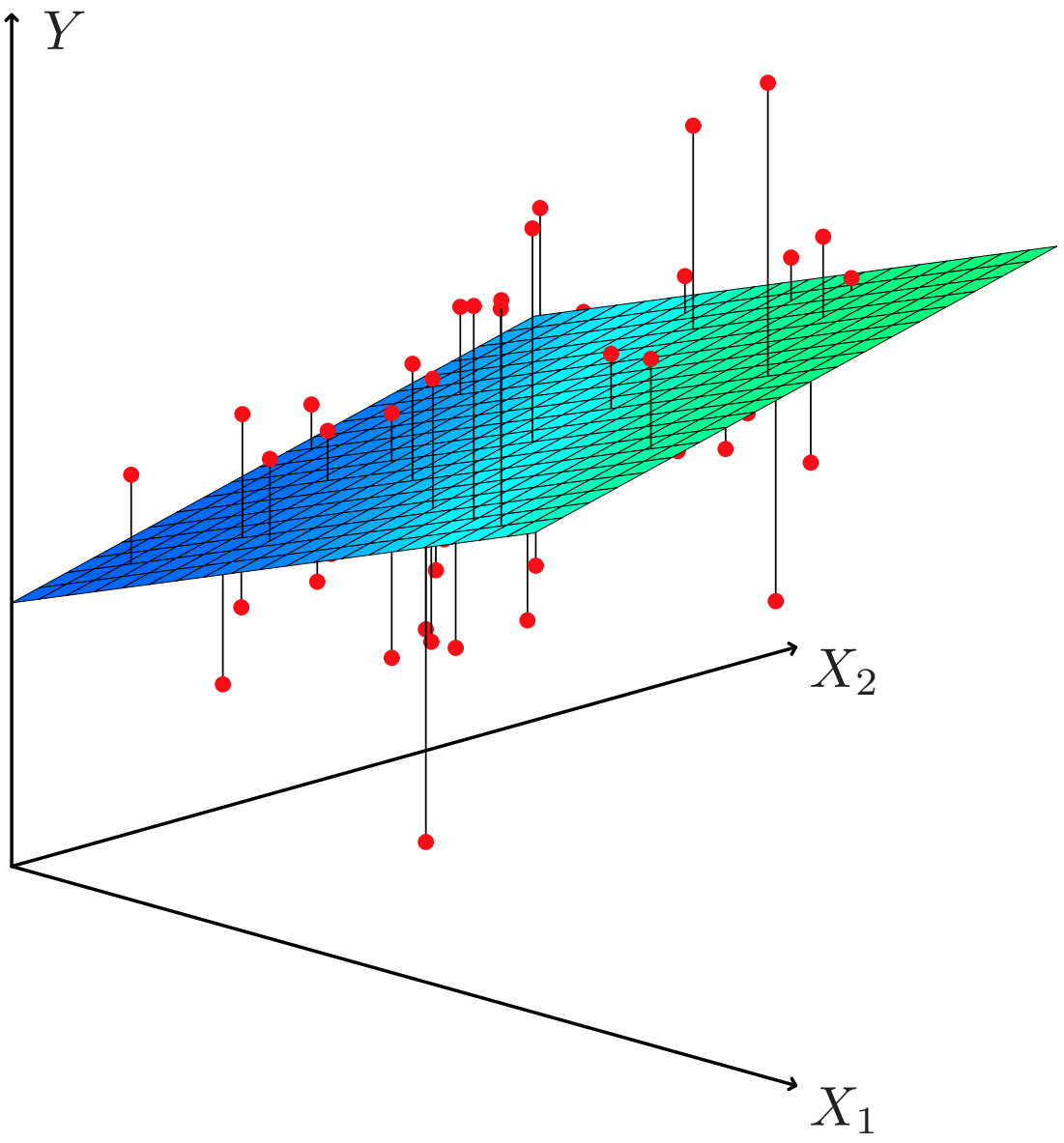
- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.


Y
X_2
X_1

Results for advertising data

Interpretation of newspaper coefficient : Sales is expected to decrease by \$1(=0.001*1000) when you increase advertising cost by \$1000(1 unit = \$1000), while holding all other predictors constant.

	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

Note: Correlation coefficient between 2 variables lies between -1 and 1 with values close to 0 indicating a weak linear relationship between the variables and values close to 1 or -1 indicating a strong positive or negative linear relationship between the variables respectively.

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

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

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

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.

Model selection — continued

- Later we discuss more systematic criteria for choosing an "optimal" member in the path of models produced by forward or backward stepwise selection.

- These include *Mallow's C_p*, *Akaike information criterion (AIC)*, *Bayesian information criterion (BIC)*, *adjusted R^2* and *Cross-validation (CV)*.

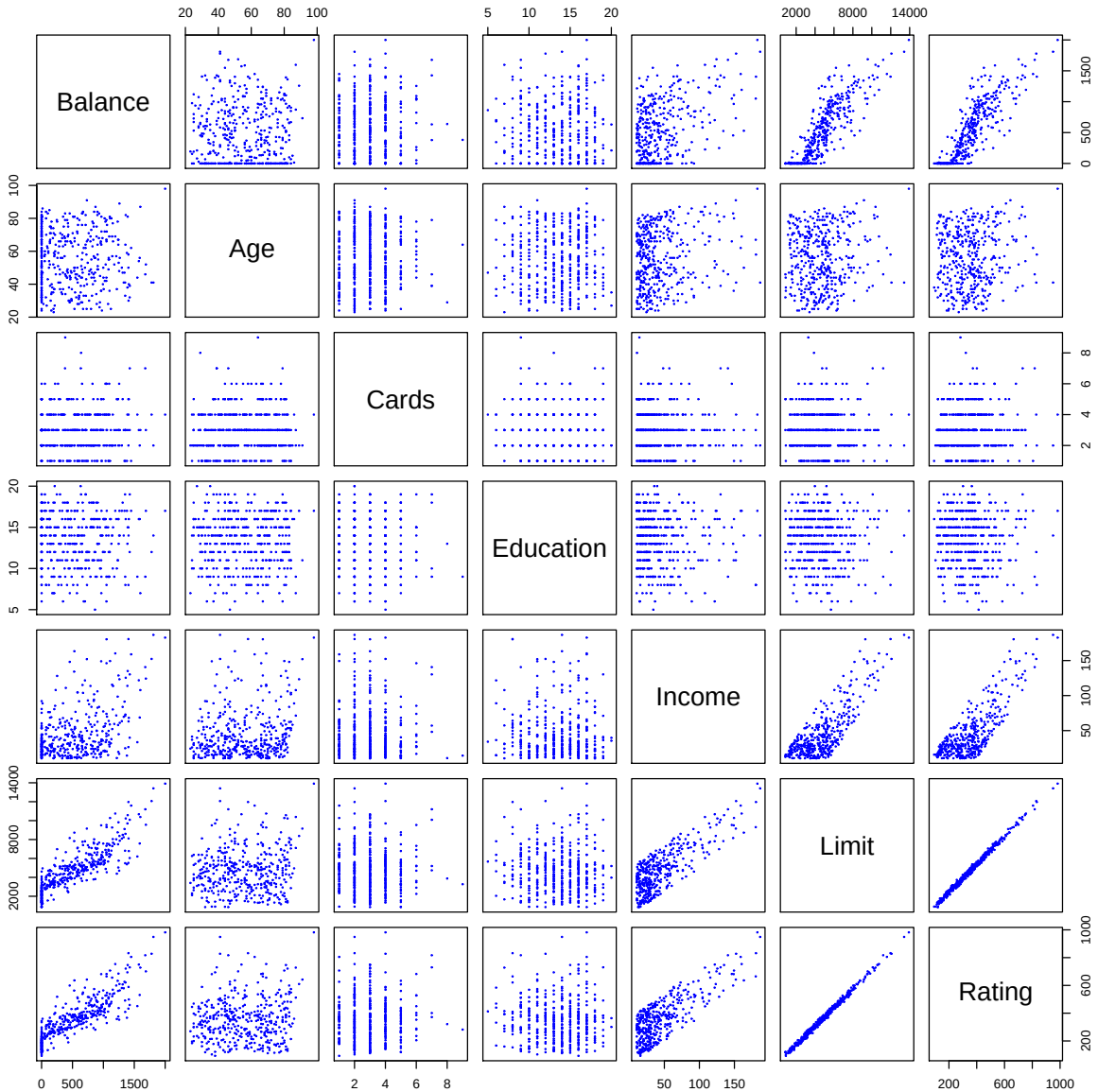
Other Considerations in the Regression Model

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*.
- See for example the scatterplot matrix of the credit card data in the next slide.

  In addition to the 7 quantitative variables shown, there are four qualitative variables: `gender`, `student` (student status), `status` (marital status), and `ethnicity` (Caucasian, African American (AA) or Asian).

Credit Card Data



Qualitative Predictors — continued

Example: investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable

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

Resulting model:

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

Intrepretation?

Credit card data — continued

Results for gender model:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

Qualitative predictors with more than two levels

- With more than two levels, we create additional dummy variables. For example, for the `ethnicity` variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

Qualitative predictors with more than two levels — continued.

- Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — African American in this example — is known as the *baseline*.

Results for ethnicity

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

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

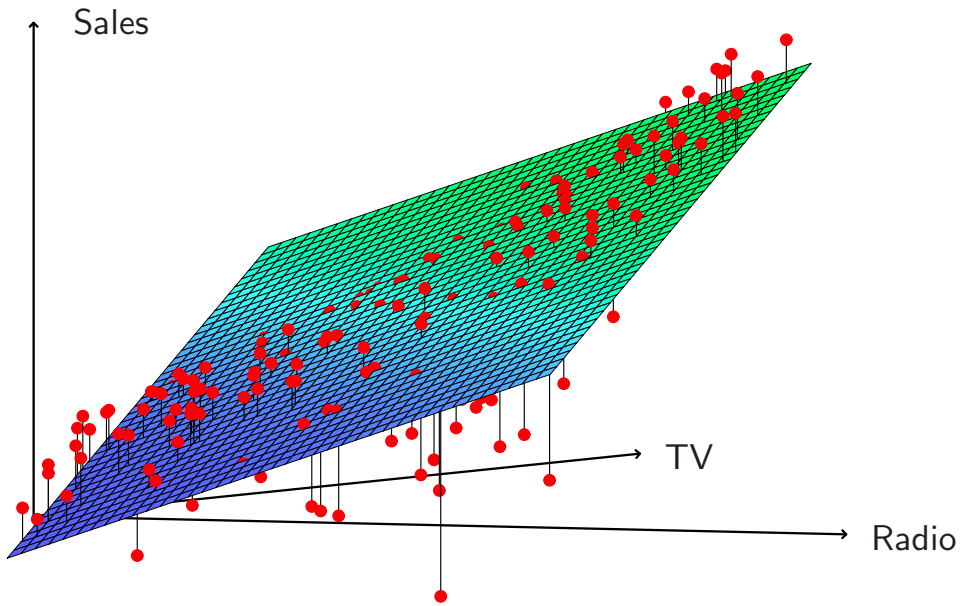
$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

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 — continued

- 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`.

Modelling interactions — Advertising data

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 $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
- An increase in radio advertising of $\$1,000$ will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

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.

Hierarchy — continued

- The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.
- Specifically, the interaction terms also contain main effects, if the model has no main effect terms.

Interactions between qualitative and quantitative variables

Consider the `Credit` data set, and suppose that we wish to predict `balance` using `income` (quantitative) and `student` (qualitative).

Without an interaction term, the model takes the form

$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}$$

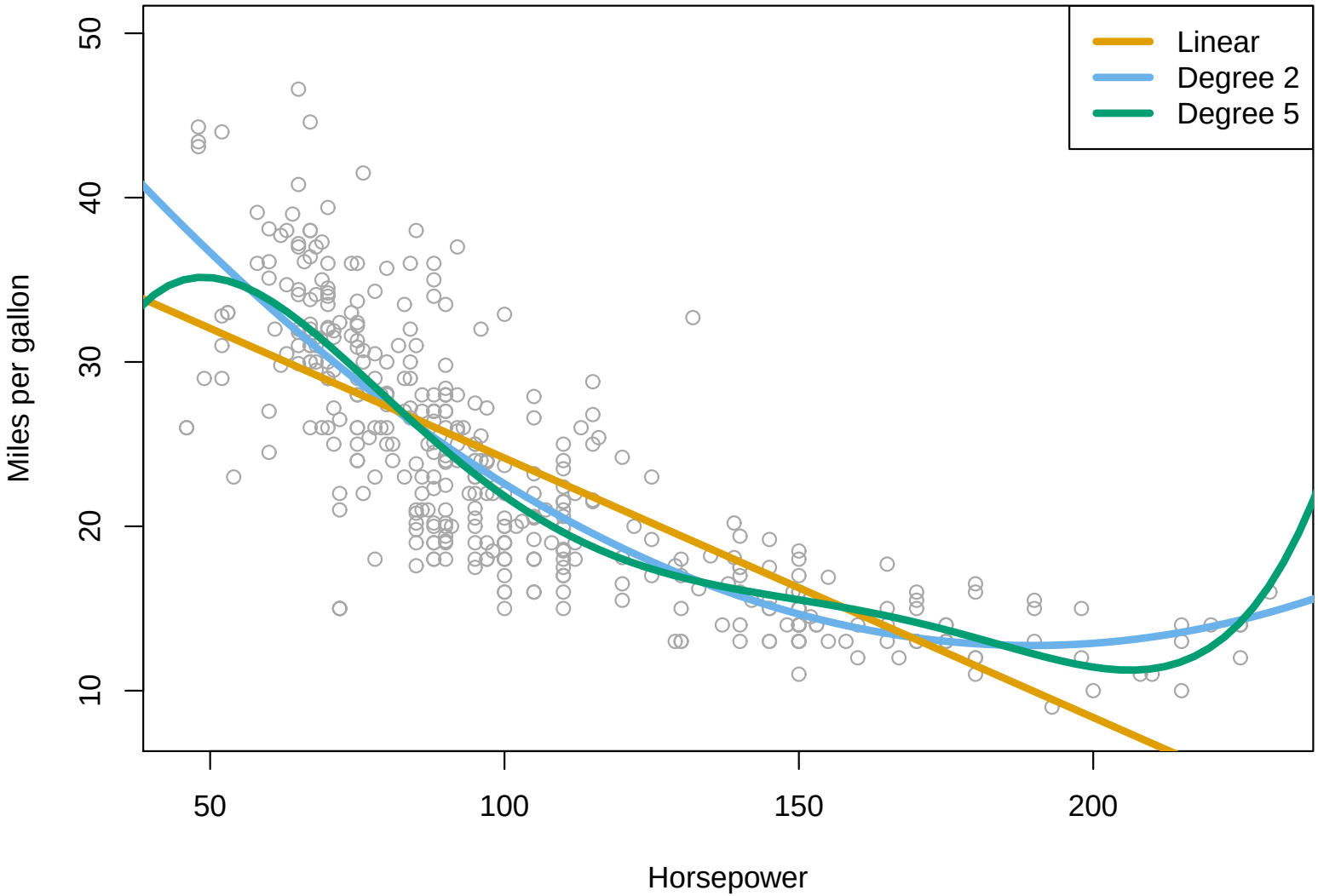
With interactions, it takes the form

$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}
\end{aligned}$$

Here, you can see that the coefficient of the interaction term ($\beta_2 + \beta_3$) is simply how much the credit card balance would increase or decrease once their income increases by \$1000 (unit income =\$1000), for a student compared to a non student.

Non-linear effects of predictors

polynomial regression on `Auto` data



The figure suggests that

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

may provide a better fit.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

Week 3 - Logistic Regression

Classification

- Qualitative variables take values in a set C, such as:

  $X_1 =$ eye color {brown, blue, green},

  $X_2 =$ hotel rating{1 star, 2 star,…, 5 star}.

- Classification is a type of supervised learning technique where the aim is to predict the categorical class labels of new instances, based on past observations

- Given a feature vector X and a qualitative response Y taking values in the set C, we use classification to build a function C(X) so that given the feature vector X, we predict its value for Y ; i.e. C(X) ∈ C.

Classification

- The output variable in classification is categorical (as opposed to continuous in linear regression). These categories are sometimes referred to as **'classes**' or **'labels'**.
- Types of Classification:

Binary Classification: The simplest form of classification where there are only two classes to predict, such as spam or not spam, sick or healthy.

Multiclass Classification: Involves categorizing data into more than two classes, like identifying the type of fruit in an image (apples, bananas, oranges, etc.).

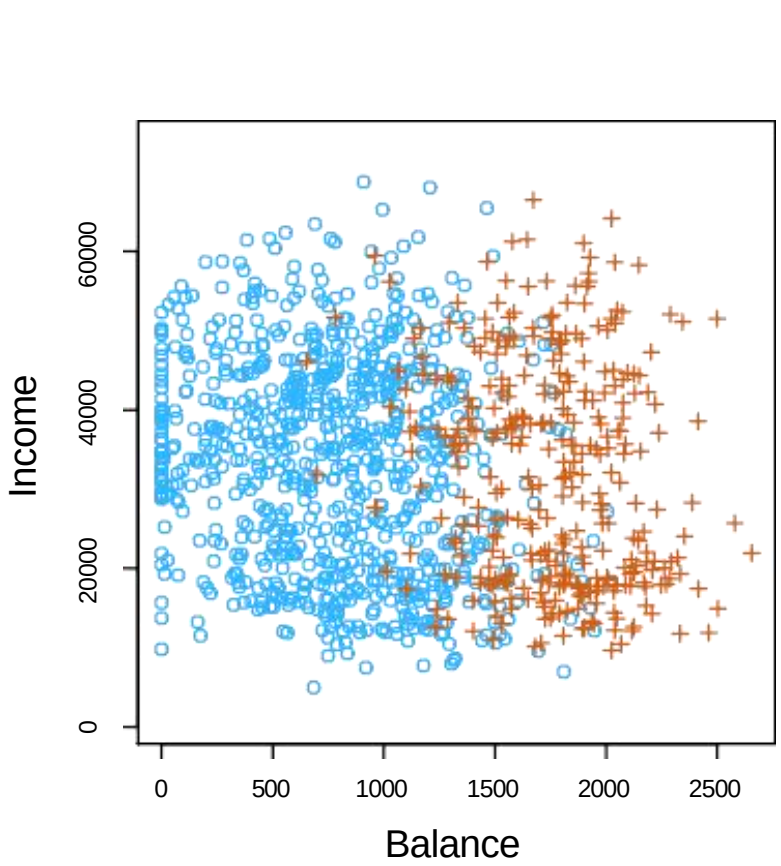
Classification

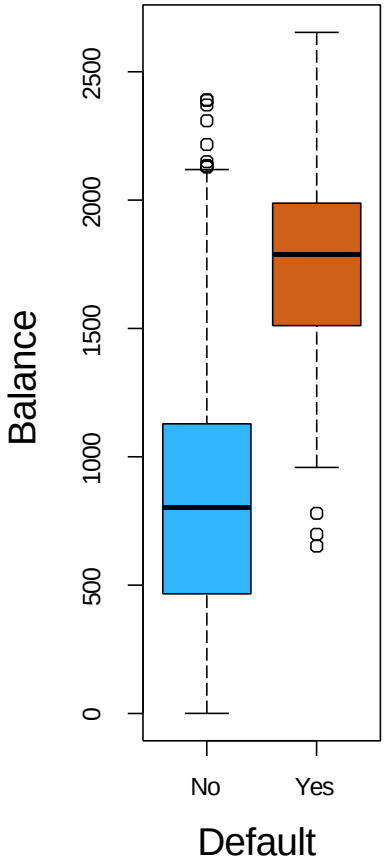
- Often we are more interested in estimating the *probabilities* that X belongs to each category in *C*.

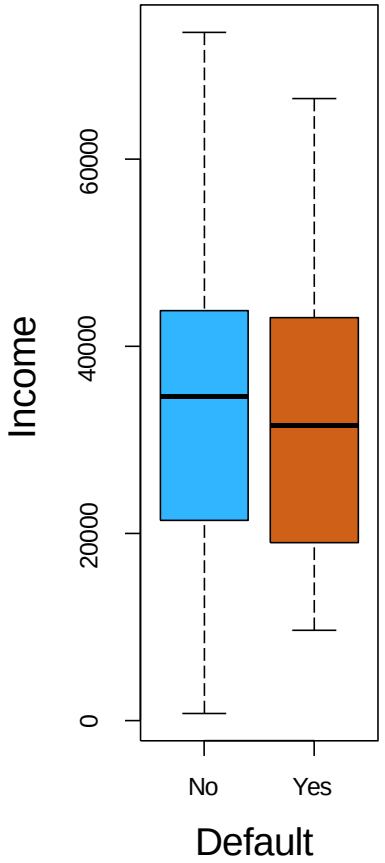
  For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

  - Logistic regression is a form of classification technique

Example: Credit Card Default







Can we use Linear Regression?

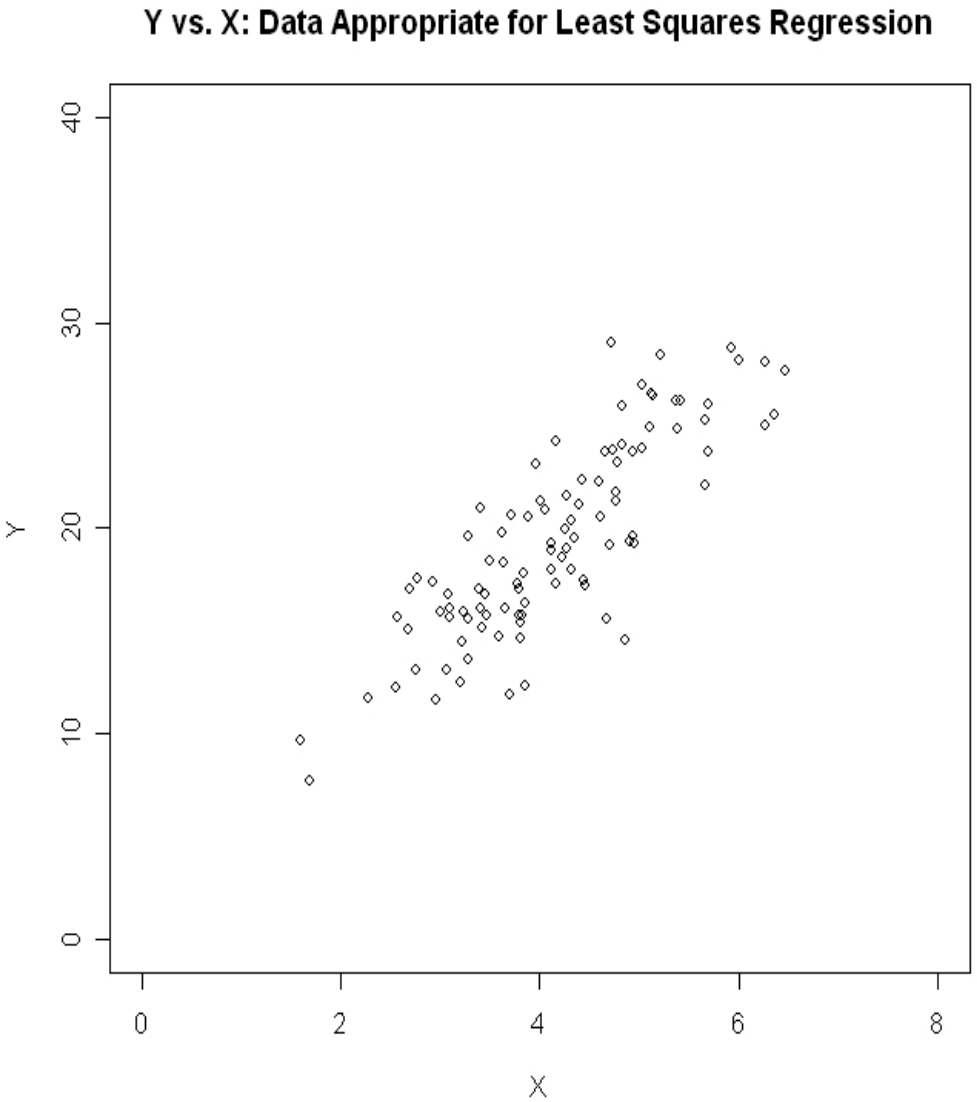
Suppose for the `Default` classification task that we code

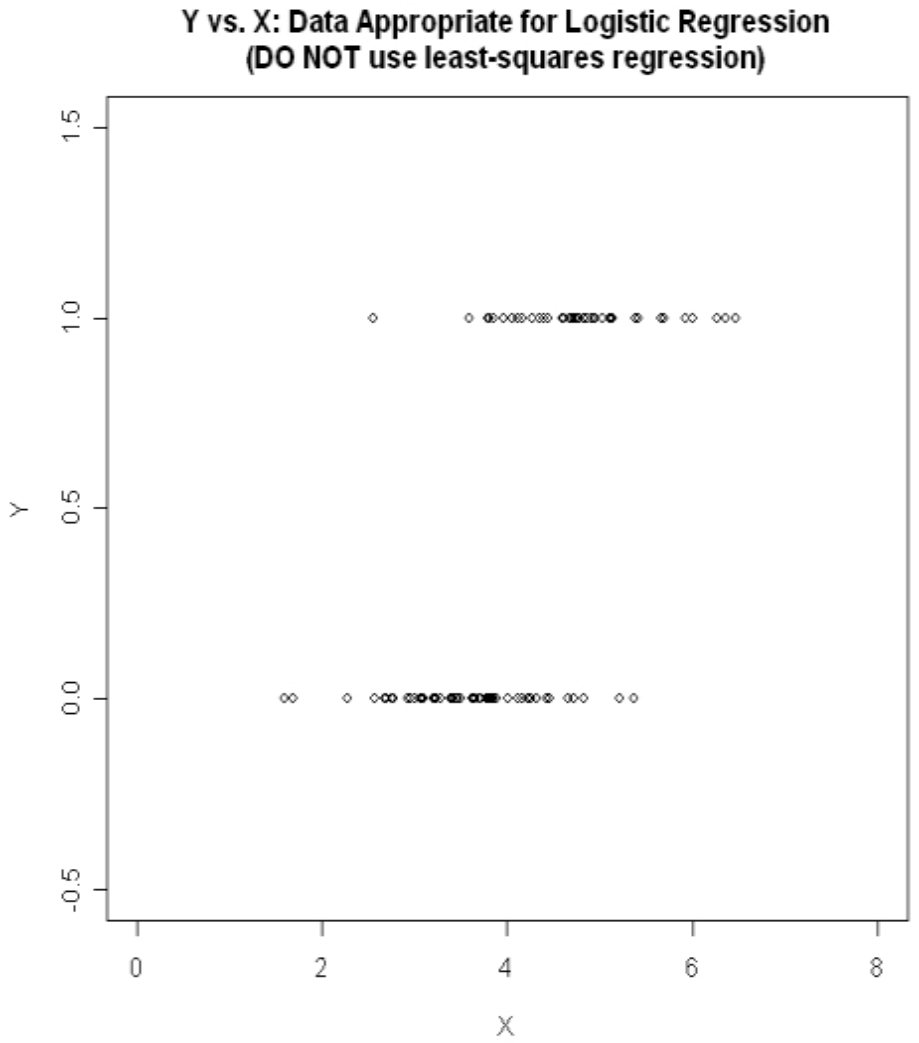
$$Y = \begin{cases} 0 & \text{if } \texttt{No} \\ 1 & \text{if } \texttt{Yes}. \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as `Yes` if $\hat{Y} > 0.5$?

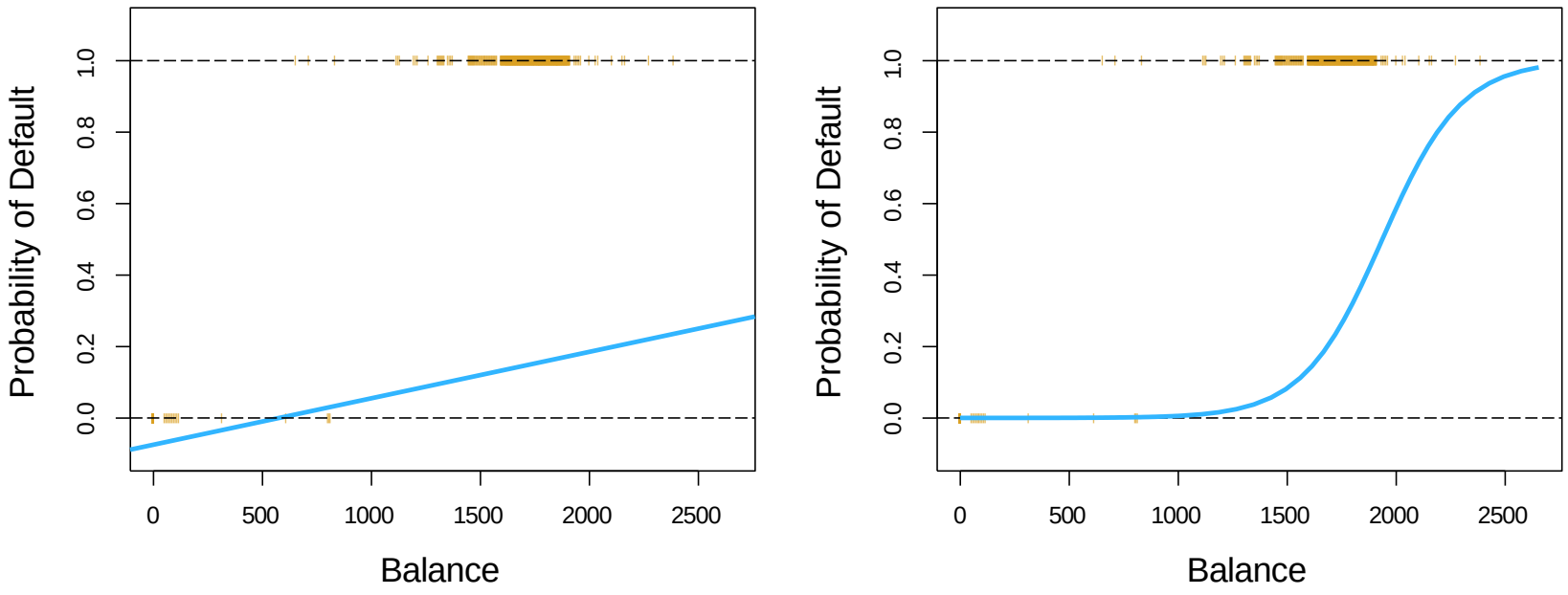
- Note that Y is a Bernoulli random variable with $E(Y) = P(Y = 1)$ and $Var(Y) = P(Y = 1)[1 - P(Y = 1)]$
- Since in the population $E(Y|X = x) = \Pr(Y = 1|X = x)$, we might think that linear regression is perfect for this task.
- However, *linear* regression might produce probabilities less than zero or bigger than one. *Logistic regression* is more appropriate.

Linear Regression vs Logistic Regression





Linear versus Logistic Regression



The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Linear Regression continued

- Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

- This coding suggests an ordering, and in fact implies that the difference between stroke and drug overdose is the same as between drug overdoseand epileptic seizure.
- Linear regression is not appropriate here. We can use *Multiclass Logistic Regression/ Multinomial Logistic Regression.*

Logistic Regression

- Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using balanceto predict default. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

($e \approx 2.71828$ is a mathematical constant [Euler's number.])

- It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

- A bit of rearrangement gives

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0+\beta_1 X$$

This monotone transformation is called the *log odds* or *logit* transformation of $p(X)$. (by log we mean *natural log*: ln.)

Maximum Likelihood

We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

This *likelihood* gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

Most statistical packages can fit linear logistic regression models by maximum likelihood. In R we use the glm function.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Making Predictions

What is our estimated probability of default for someone with a balance of $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

With a balance of $2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

Lets do it again, using studentas the predictor.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	$<$ 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Logistic regression with several variables

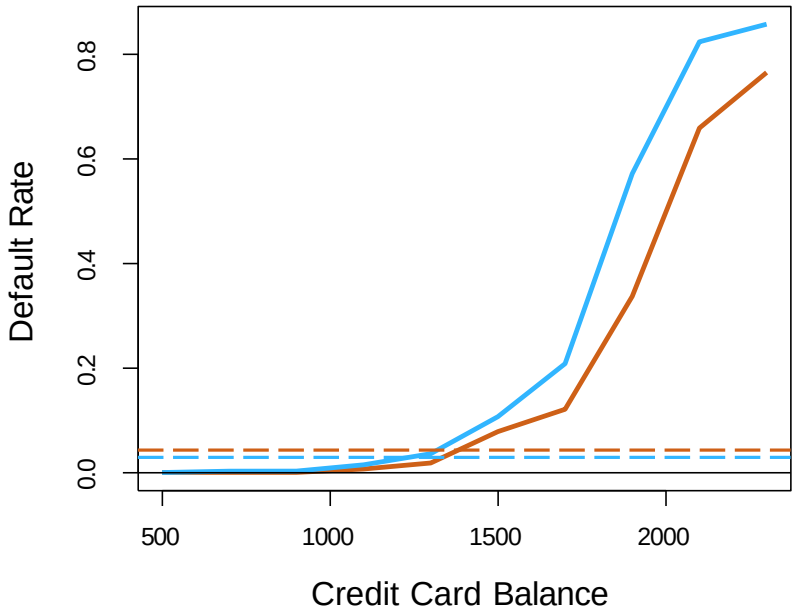
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

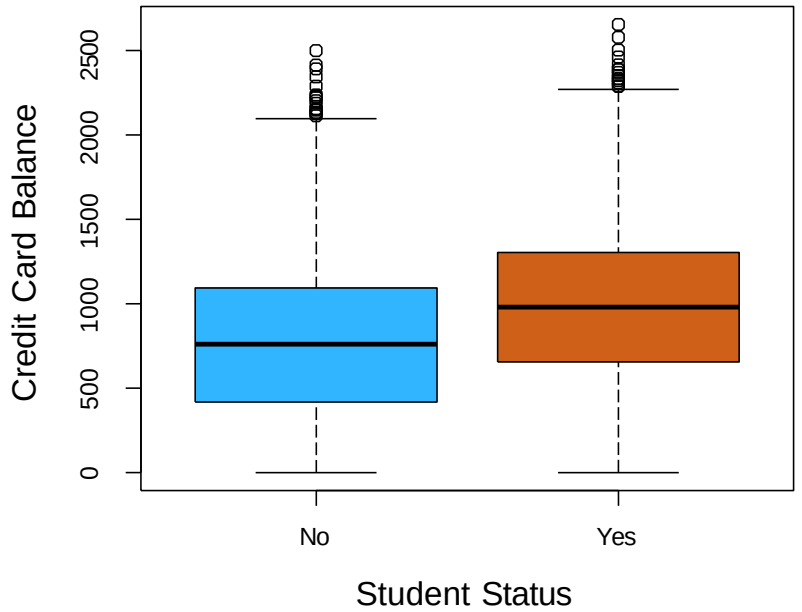
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Why is coefficient for student negative, while it was positive before?

Confounding





- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- We can use multiple logistic regression in this case where we have more than one predictor variable.

Example 2: Programming Task

- A systems analyst studied the effect of computer programming experience on ability to complete within a specified time a complex programming task, including debugging.
- Twenty-five persons were selected for the study.
- $X_1 =$ Months of programming experience (measured in months of experience - column 1)
- All persons were given the same programming task, and the results of their success in the task are shown in column 2 of the data.
- The results are coded in binary fashion: Y = 1 if the task was completed successfully in the allotted time, and Y = 0 if the task was not completed successfully.

Example 2: Programming Task

- From the graph, we see that the probability of success increases sharply with experience.

Example 2: Programming Task

- The estimated logistic regression function is given as:

$$\hat{P}(Y = 1|X) = \frac{\exp(-3.059696 \ +0.16149X)}{1 + \exp(-3.059696 \ + \ 0.16149X)}$$

Or

$$\log\left(\frac{\hat{P}(Y = 1|X)}{1 - \hat{P}(Y = 1|X)}\right) = -3.059696 \ +0.16149X$$

- What is the estimated probability that a task will be completed by a person who has 14 months of experience?

$$\hat{P}(Y = 1|X = 14) = \frac{\exp(-3.059696 \ +0.16149(14))}{1 + \exp(-3.059696 \ + \ 0.16149(14))}$$
$$= \ 0.3013$$

Interpretation of $\hat{\beta}_1$

- The interpretation of the estimated regression coefficient $\hat{\beta}_1$ in the fitted logistic response function is not the straightforward interpretation of the slope in a linear regression model.

- An interpretation of $\hat{\beta}_1$ is found in the property of the fitted logistic function that the estimated odds $\left(\frac{\hat{P}(Y=1|X)}{1-\hat{P}(Y=1|X)}\right)$ are **multiplied by** exp $(\hat{\beta}_1)$ for any unit increase in X (need to keep all other variables constant in the case of a multiple logistic regression).

- An alternative way of interpreting $\hat{\beta}_1$ is that the estimated log odds of "success" (Y=1) changes (**increases or decreases**) by $\hat{\beta}_1$ units for a unit increase in X (need to keep all other variables constant in the case of a multiple logistic regression).

Example: Programming Task

- For the programming task example, we see from the scatter plot on page 16 that the probability of success increases sharply with experience.

- Specifically, the model's results shows that the odds ratio is

$$\text{OR} = \exp(\hat{\beta}_1) = \exp(.1615) = 1.175,$$

so we can say that the estimated odds of completing the programming task increase by 17.5 percent with each additional month of experience.(same as saying that the estimated odds of completing the programming task increase by a factor of 1.175 with each additional month of experience)

- Since a unit increase of one month is quite small, the estimated odds ratio of 1.175 may not adequately show the change in odds for a longer difference in time.
- In general, the estimated odds ratio when there is a difference of c units of X is $\exp(c\,\hat{\beta}_1)$.
  - ✓ For example, should we wish to compare individuals with relatively little experience to those with extensive experience, say 10 months versus 25 months so that c = 15, then the odds ratio would be estimated to be exp[15(.1615)] = 11.3.
  - ✓ This indicates that the odds of completing the task increase over 11-fold for experienced persons compared to relatively inexperienced persons.

Example 3: Disease outbreak

- In a health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes, individuals were randomly sampled within two sectors in a city to determine if the person had recently contracted the disease under study.

- This was ascertained by the interviewer, who asked pertinent questions to assess whether certain specific symptoms associated with the disease were present during the specified period.

- The response variable Y was coded 1 if this disease was determined to have been present, and 0 if not.

- Three predictor variables were included in the study, representing known or potential risk factors. They are age, socioeconomic status of household, and sector within city.
  - ✓ Age (X_1) is a quantitative variable.
  - ✓ Socioeconomic status is a categorical variable with two levels. It is represented by one indicator variable; $X_2 = 0$ if low and $X_2 = 1$ if high
  - ✓ City sector is a categorical variable with two levels. It is represented by one indicator variable; a $X_3 = 0$ for sector 1 and $X_3 = 1$ for sector 2.

Example 3: Disease outbreak

- The interpretation of the estimated regression coefficients in the fitted multiple logistic response function, is similar to that for the simple logistic response function

- The only difference in interpretation for multiple logistic regression is that the estimated odds ratio for predictor variable X_1 assumes that all other predictor variables are held constant : $\exp(\hat{\beta}_k)$ is the estimated odds ratio for predictor variable X_k, while holding all other predictors constant.

- For instance, the odds of a person having contracted the disease increase by a factor of $\exp(\hat{\beta}_1) = \exp(0.01402) = 1.01412$ with each additional year of age (X_1), for any given socioeconomic status and city sector location.

Example 3: Disease outbreak

- With a coefficient of 0.06964 for X_2 (Socioeconomic Status), the odds of contracting the disease for individuals with high socioeconomic status are approximately exp(0.06964)=1.072 times the odds for those with low socioeconomic status. This indicates a slight increase in risk for individuals with high socioeconomic status compared to those with low socioeconomic status.

- Since the coefficient for X_3 (City Sector) is −1.16093, this would mean that living in sector 1 of the city is associated with a decrease in the odds of contracting the disease compared to living in sector 0. Hence, the odds of contracting the disease for individuals in sector 2 are approximately exp(−1.16093) = 0.313 times the odds for those in sector 1. This indicates that individuals in sector 1 have significantly lower odds, or a lower risk, of contracting the disease compared to individuals in sector 0

- Intercept = −0.41058, meaning the odds of contracting the disease for an individual with average age, low socioeconomic status, and living in sector 0 of the city would be approximately 0.663 (baseline).

Logistic regression with more than two classes

- It is possible to extend the two-class logistic regression approach to the setting of K > 2 classes.
- Multiclass logistic regression is also referred to as *multinomial regression*.
- To do this, we first select a single multinomial logistic regression class to serve as the *baseline*;
- Without loss of generality, we can select the Kth class as the baseline.

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

for $k = 1, \ldots, K-1$, and

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

It is not hard to show that for $k = 1, \ldots, K-1$,

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p.$$

Logistic regression with more than two classes

- The choice of baselines does not matter.
- The coefficient estimates will differ between any two models fitted with different choices of baselines but the fitted values (predictions), the log odds between any pair of classes, and the other key model outputs will remain the same

Logistic regression with more than two classes

- Interpretation of the coefficients in a multinomial logistic regression model is tied to the choice of baseline.
- Refer to the example on page 8. Suppose we set epileptic seizure to be the baseline, we can interpret the intercept $\hat{\beta}_{stroke0}$ as the log odds of stroke versus epileptic seizure, given that $X_1 = \cdots = X_p = 0$.
- Furthermore, a one-unit increase in X_j is associated with a $\hat{\beta}_{strokej}$ increase in the log odds of stroke over epileptic seizure.
- Alternatively, we can say that, if X_j increases by one unit, then the odds of getting a stroke vs an epileptic seizure, that is

$$\frac{\Pr(Y = \texttt{stroke}|X = x)}{\Pr(Y = \texttt{epileptic seizure}|X = x)}$$

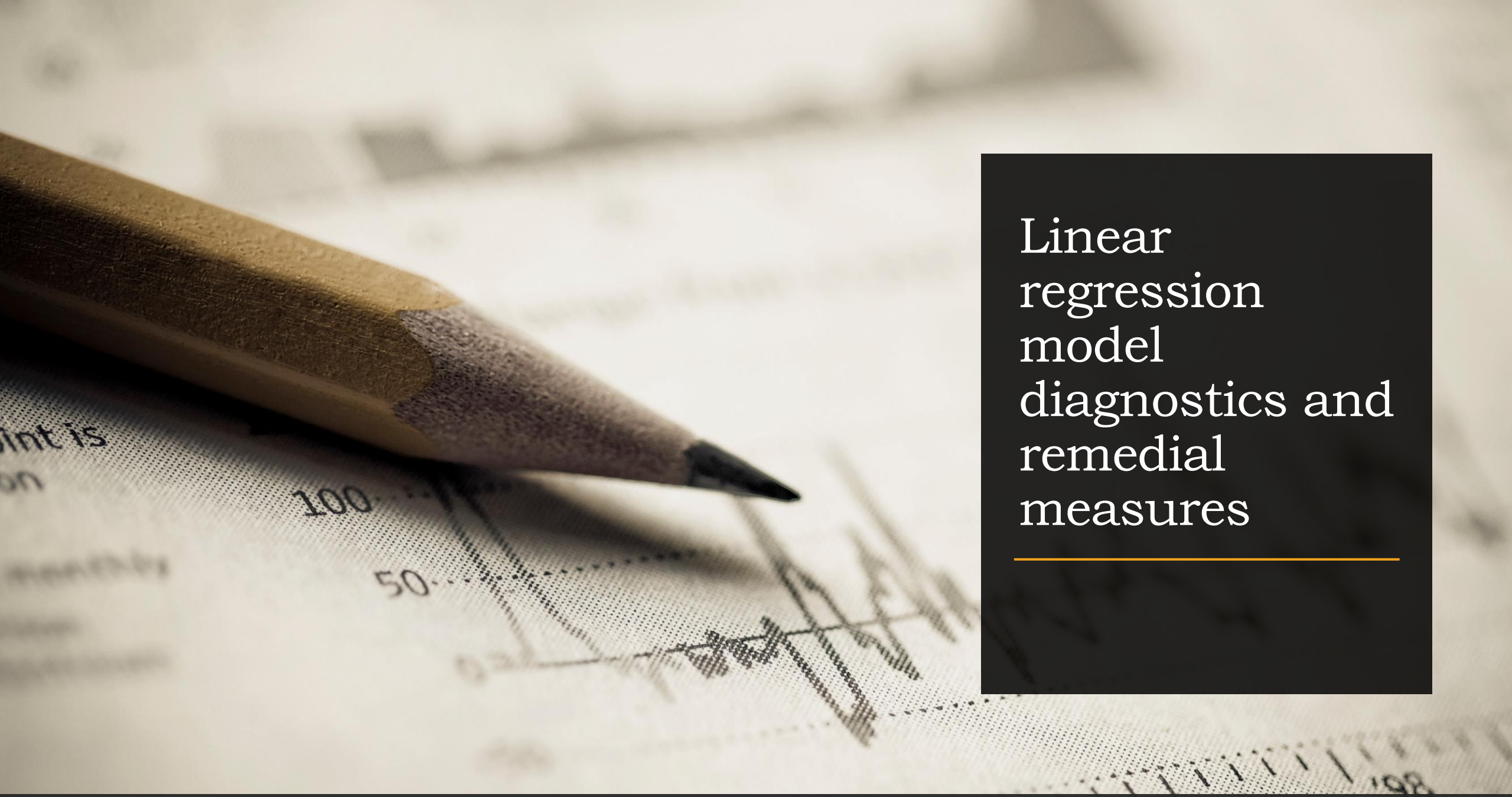
increases by $e^{\widehat{\beta}_{strokej}}$.

REFER TO R FOR THE IRIS DATA EXAMPLE – EXAMPLE 4

UP NEXT: Model Diagnostics and Validation

References

Hastie T, Tibshirani R, Friedman J. An introduction to statistical learning.
Neter, John, et al. "Applied linear statistical models." (1996): 318.



Linear regression model diagnostics and remedial measures

Diagnostics for Linear Regression

- It is important to examine the appropriateness of the LR model for the given data before inferences and predictions based on that model are undertaken.
  - This is done by checking model assumptions.
- Violations in these assumptions lead to inaccurate conclusions, predictions and estimates.
- Once the linear model has been fit, model diagnostics need to be performed before any inferences are made from the model.
- We can do this using graphical methods or formal statistical tests.

Model assumptions to check

- Diagnostics for the model are usually carried out indirectly by examining the residuals, $e_i = Y_i - \hat{Y}_i$.
- The assumptions on the errors are
  - Independent, normally distributed, constant variance, ie $\varepsilon_i \sim iid\ N(0, \sigma^2)$
- Another assumption is the linearity between X and Y.
- With multiple predictors, the relationships between predictors must also be considered.

Residual Diagnostics

- The following questions must be addressed by diagnostics for residuals
  1. Is the relationship linear?
  2. Is the variance constant or does it depend on X?
  3. Are the errors normal?
  4. Are the errors independent?
  5. Are there outliers?
  6. Can other predictors be helpful?

Residual Plots

- Used to assess the assumption on the errors:
  - Linearity, Independence, Normality, Constant variance (LINC)
- Normality (Normal probability plot)
- Residuals vs. Fitted Values (Constant Variance)
- Residuals vs. Predictor Variables (Constant Variance, Linearity)
- Residuals vs. Order (Independence, if applicable)
- Residuals vs. Other Variables (e.g., interaction terms) can help determine if omitted variables may be important to include in the model.

Scatterplot of Y vs. X - SLR

- The scatterplot is a common diagnostic to view the nature of the relationship between X and Y.
- It is useful for
  - Checking to see if a linear trend is reasonable.
  - Visualizing the strength and direction of the relationship.
  - Checking for unusual observations (outliers) for X or Y.
  - Checking the scope of the model by looking at the range of the X values.

Scatterplot Matrix - MLR

- Summarizes bivariate relationships:
  - Between response and explanatory variables
  - Among explanatory variables(multicollinearity)
- Serves as an initial diagnostic tool for:
  - Assessing the nature and strength of bivariate relationships
  - Detecting outliers
  - Observing the range covered by the variables

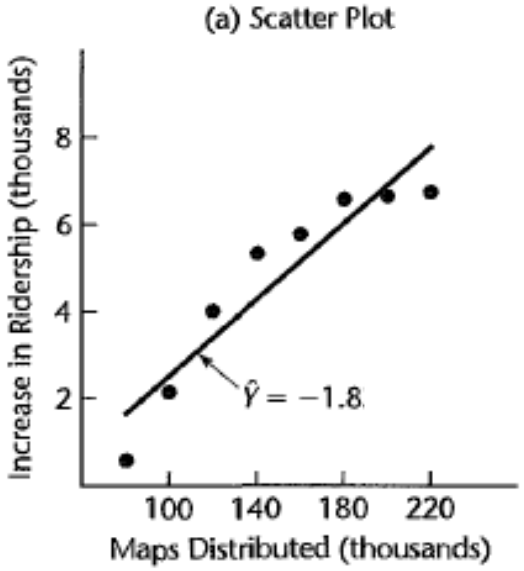
Correlation Matrix - MLR

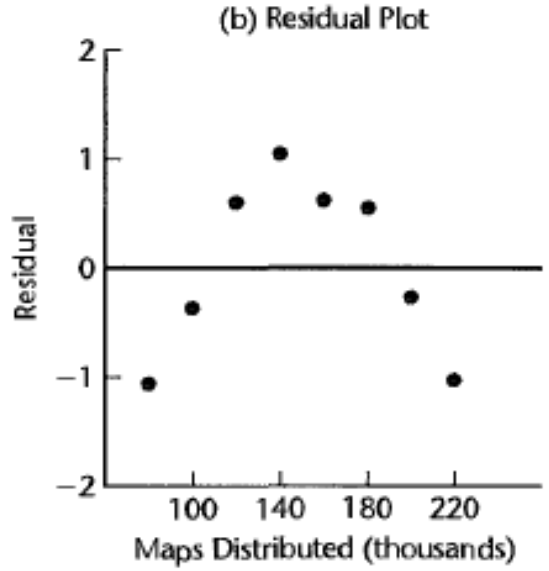
- Can be used in addition to the scatterplot matrix
- Gives all pairwise correlations - the strength and direction of linear relationships between pairs of variables.
- Identifies potential multicollinearity issues among explanatory variables .

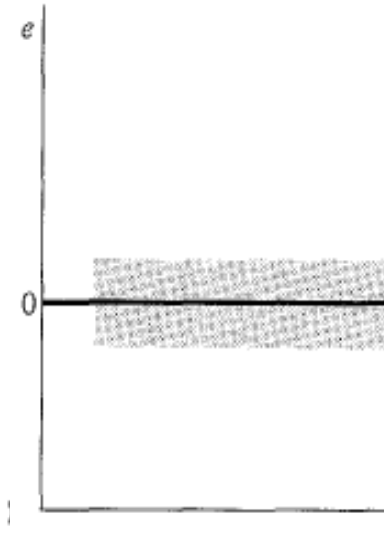
Checking linearity

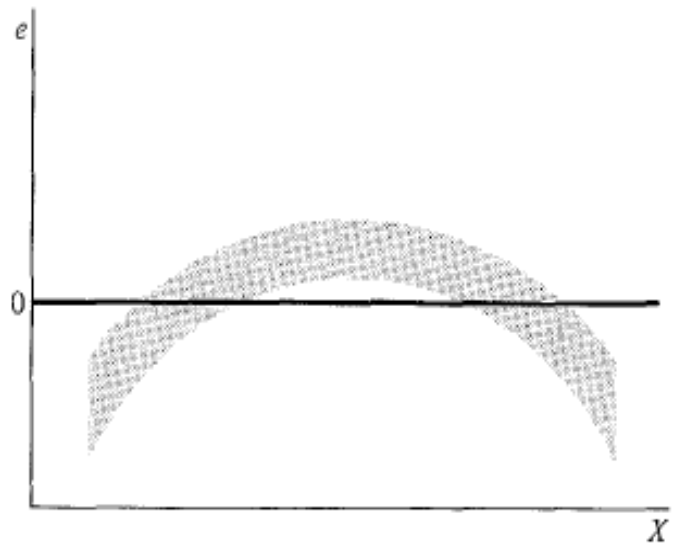
1. Plot of residuals, e against fitted values $\hat{Y}$
2. Plot of residuals, e against each predictor variable, X (gives same information as plot in 1)
3. Plot of Y vs each X (scatterplot- this is not as effective as a residual plot as seen below)

- The residuals should fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative.





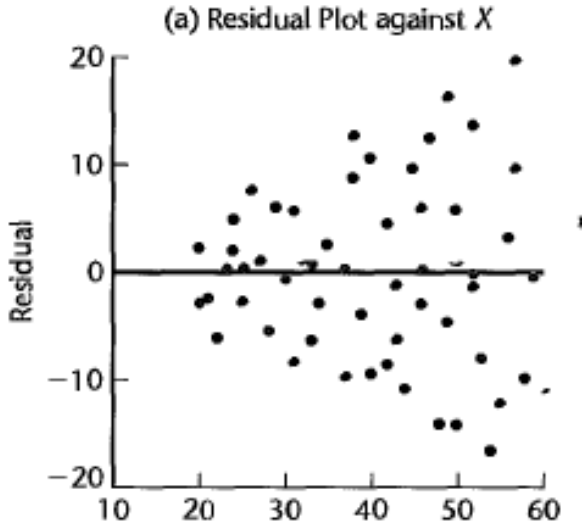


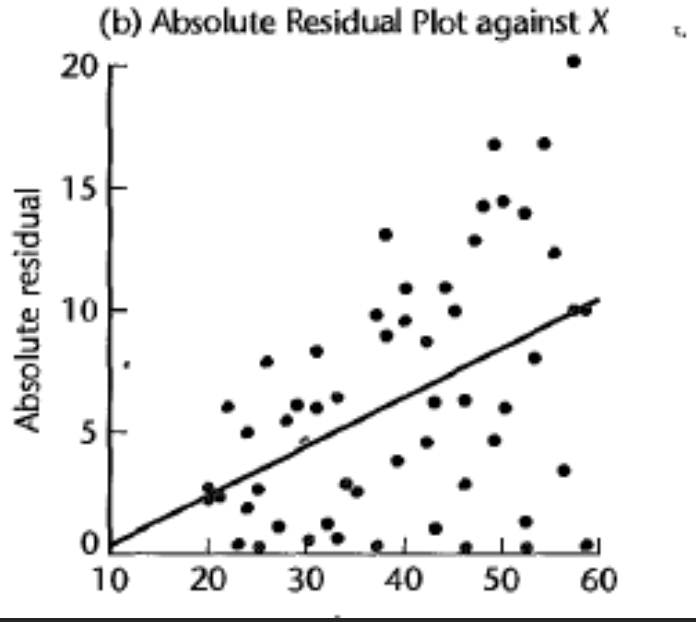


Checking constant variance(homoscedasticity)

1. Plot of residuals against fitted values.
2. Plot of residuals against each predictor variable, X.
3. Plots of the absolute values of the residuals or of the squared residuals against the predictor variable X or against the fitted values $\hat{Y}$

- A random scattering of points around zero indicates the assumptions are met (this is desired)
- A megaphone shape indicates error variances decreasing with increasing levels of the predictor variable.





Residual plots indicating non constant variance

Checking normality

1. Normal probability plot (Q-Q plot) of residuals, e
   - Points should hug the Q-Q line tightly.
   - Plot should be symmetric; heavy tails show the ends moving away from the Q-Q line (in an S shape).
2. Histogram of residuals, e .
   - Histograms should be symmetric if normal

Checking for presence of outliers

- Outliers are extreme observations.
- Residual outliers can be identified from residual plots against X or $\hat{Y}$, as well as from box plots, stem-and-leaf plots, and dot plots of the residuals.

Checking independence

1. Plot of residuals, *e* against time or other sequence.
   - This should be constructed when the data are obtained in a time sequence or some other type of order.
   - We expect residuals to be scattered randomly around zero if errors are independent
   - Patterns (cyclical, positive or negative trend) suggest non-independence

Residuals against additional predictors

1. Plot residuals (*e*) vs. other potential predictors
2. Patterns indicate an important predictor that should be in the model

Summary of Diagnostic Plots

- Several of the same plot can be used to check multiple assumptions.

1. Plot of Y against X
   - Check linearity, outliers
2. Plot of residuals against X
   - Check for linearity, constant variance, outliers
3. Normal probability plot or histogram of residuals
   - Check normality, outliers
4. Consider a sequence plot (residuals vs. order/ time) to check for independence, i.e. if data was collected in order.
5. If there are any other predictors that are thought to influence the response you can also plot (residuals vs. new predictor).

Remedial measures to assumption violation

Non-linearity

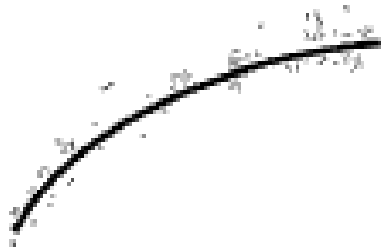
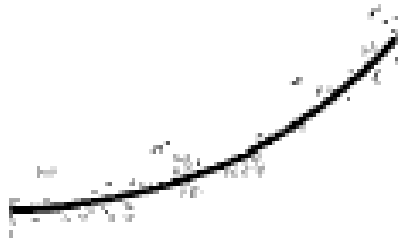
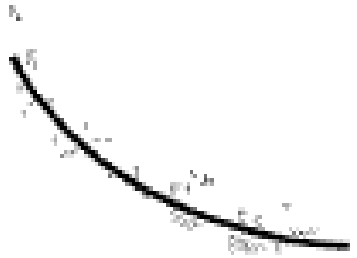
- Remedy: Transformation of X
- Look at scatterplots and residual vs. X plot for patterns. These plots also give an idea of the appropriate transformation procedure to use.
- We transform X when the error assumptions are satisfied yet there is the problem of nonlinearity.
- Can potentially still use a "linear" model:
  - For Example

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 \ln X + \varepsilon$$

  - The model is still "linear" in terms of the regression coefficients (parameters)
- To transform X, simply consider a new predictor variable X^2 or $\ln X$ and treat like any usual predictor.
- Another remedy is using nonlinear regression

Some general patterns:

Prototype Regression Pattern		Transformations of X	
(a)		$X' = \log_{10} X$	$X' = \sqrt{X}$
(b)		$X' = X^2$	$X' = \exp(X)$
(c)		$X' = 1/X$	$X' = \exp(-X)$

Non-constant variance

- Unequal error variances and nonnormality of the error terms frequently appear together.
- Remedy: Transformation of Y often solves both problems
- Some common transformations are
  - $\ln Y$- decreasing variance
  - $1/Y$- increasing variance
  - $\sqrt{Y}$ - increasing variance
- It is often difficult to determine from diagnostic plots, which transformation of Y is most appropriate for correcting
- The Box Cox procedure can be used. (More on this later)
- Apply the transformation, fit the model again, and re-check assumptions to see if transformation helped.
- Alternative approach is Weighted Least Squares

Non-normality of errors

- If the distribution of the error terms is known, we can use generalized linear model
  - Examples of the response distribution are:
    - Binomial (for binary categorical responses)
    - Poisson (response is a count)
- If the distribution of the error terms is unknown, we can use transformation of the response, Y
- It is often difficult to determine from diagnostic plots, which transformation of Y is most appropriate for correcting

Box Cox transformation

- The **Box Cox** procedure automatically identifies a transformation from the family of power transformations on Y.
  - It can be used to get an idea of λ
- The family of power transformations on Y is of the form

$$Y' = Y^{\lambda} \text{ to fit the model}$$

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- To estimate , we use
  - Maximum likelihood estimation
  - Numerical search procedure
- Transformations have a capable of stabilizing both variance and improve normality

Some general transformations:

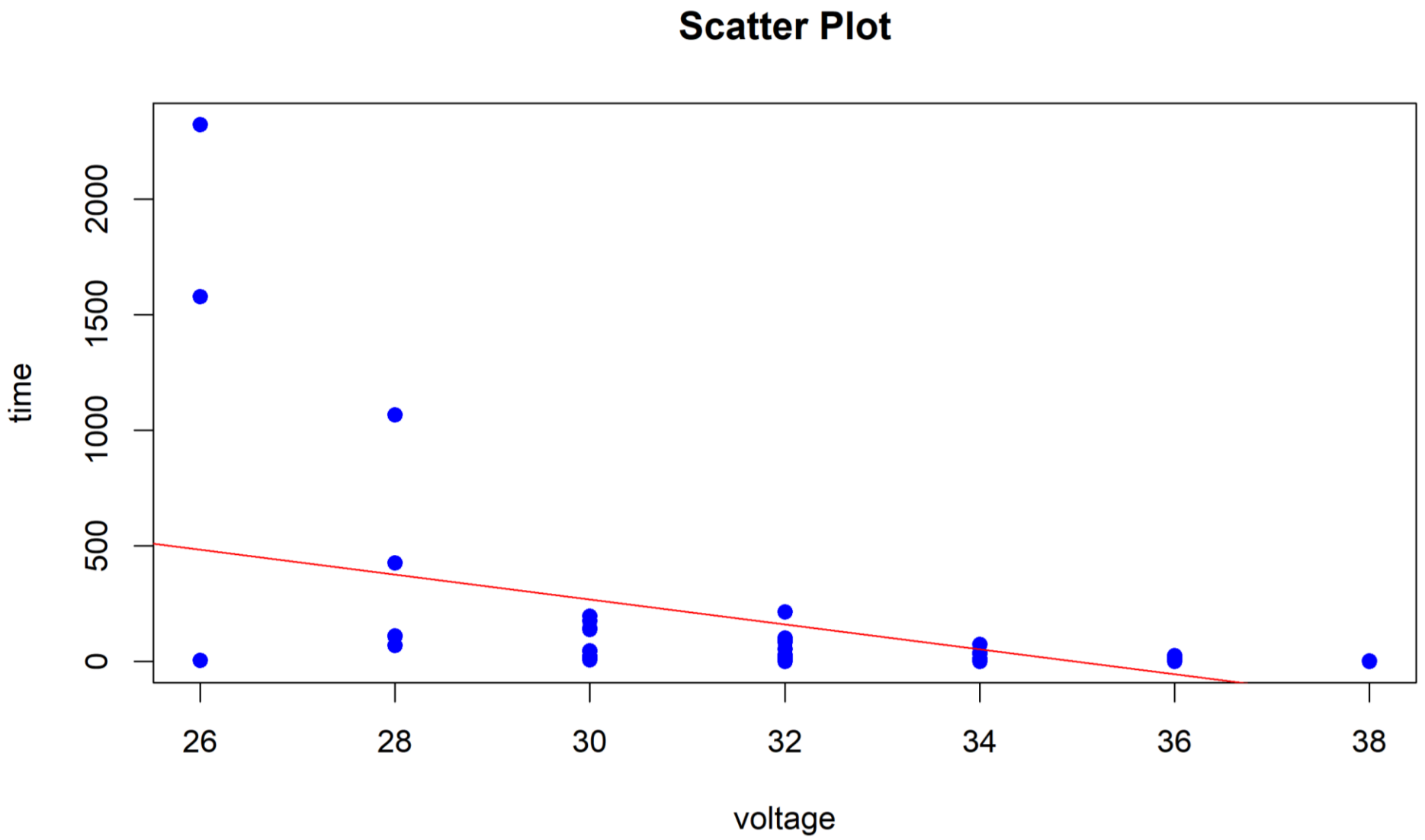
λ	Transformation
-2	$\frac{1}{Y^2}$
-1	$\frac{1}{Y}$
-0.5	$\frac{1}{\sqrt{Y}}$
0	$\ln Y$
0.5	$\sqrt{Y}$
1	Y
2	Y^2

Summary of remedial measures

1. Non- linear relationship
   - Transform X
   - Use non linear regression methods
2. Non- constant variance
   - Transform Y
   - Use weighted least squares approach
3. Non- normality
   - Transform Y
   - Use Generalized Linear Models
4. Non- independence
   - Allow correlated errors in the model

Insulating Fluid example:

- Batches of electrical insulating fluid were subjected to constant voltages until the insulating property of the fluids broke down
- Seven different voltage levels spaced 2 kilovolts (kV) apart from 26 to 38 kV were studied.
- The response measured was time (in minutes) until breakdown
- Investigate the relationship between: Voltage (X) and Breakdown Time (Y)
- Data file: insulatingfluid.csv
- R code: Week 4

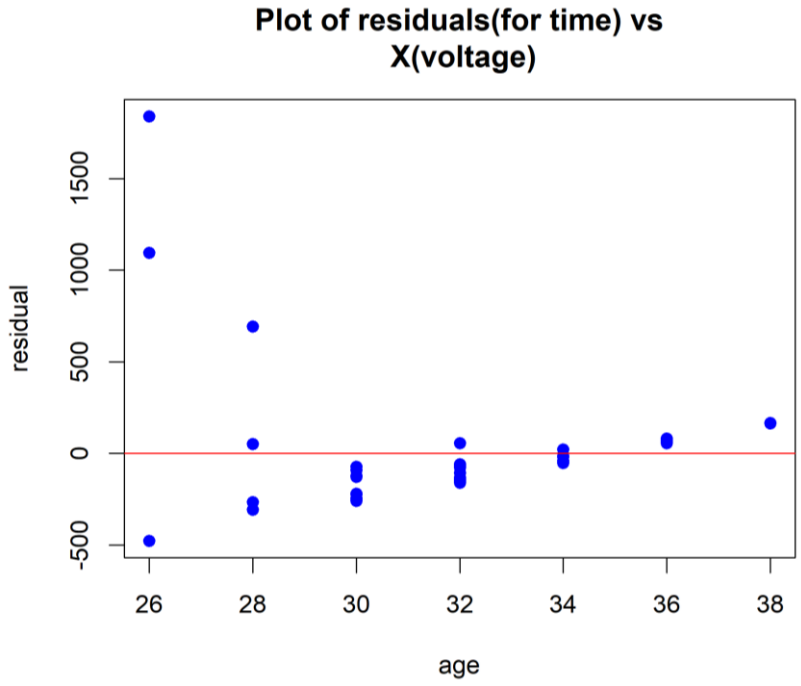


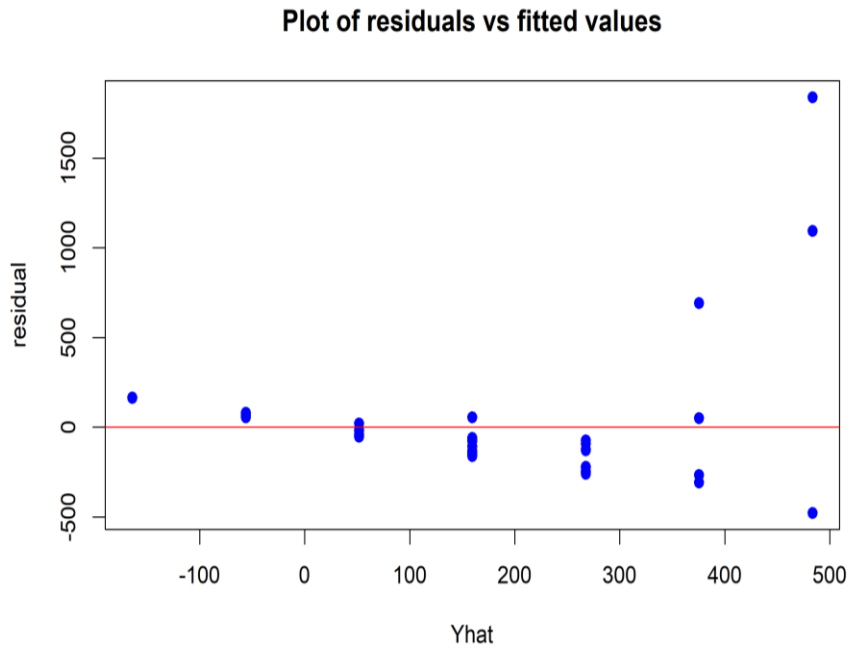

```
> cor(time,voltage)
[1] -0.4969439
```

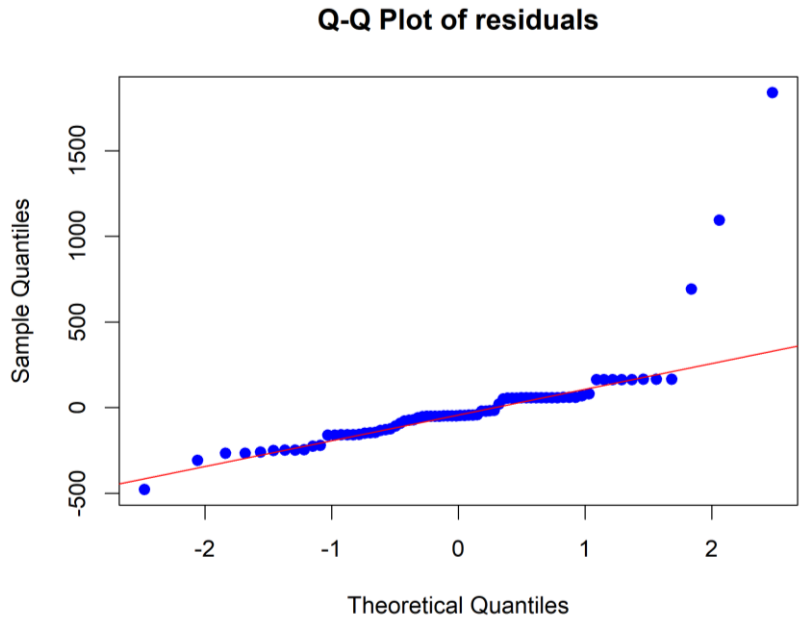
- Trend does not appear to be a strong linear trend
- There is more variation for smaller voltages than large voltages

```
> summary(model)

Call:
lm(formula = time ~ voltage)

Residuals:
    Min      1Q  Median      3Q     Max
-477.55 -144.18  -46.03   58.41 1840.36

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1886.17     364.48   5.175 1.89e-06 ***
voltage       -53.95      10.95  -4.926 4.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 297.7 on 74 degrees of freedom
Multiple R-squared:  0.247,     Adjusted R-squared:  0.2368
F-statistic: 24.27 on 1 and 74 DF,  p-value: 4.966e-06

> anova(model)
Analysis of Variance Table

Response: time
          Df  Sum Sq Mean Sq F value    Pr(>F)
voltage    1 2150408 2150408  24.267 4.966e-06 ***
Residuals 74 6557345   88613
---
```

Diagnostics of the SLR model

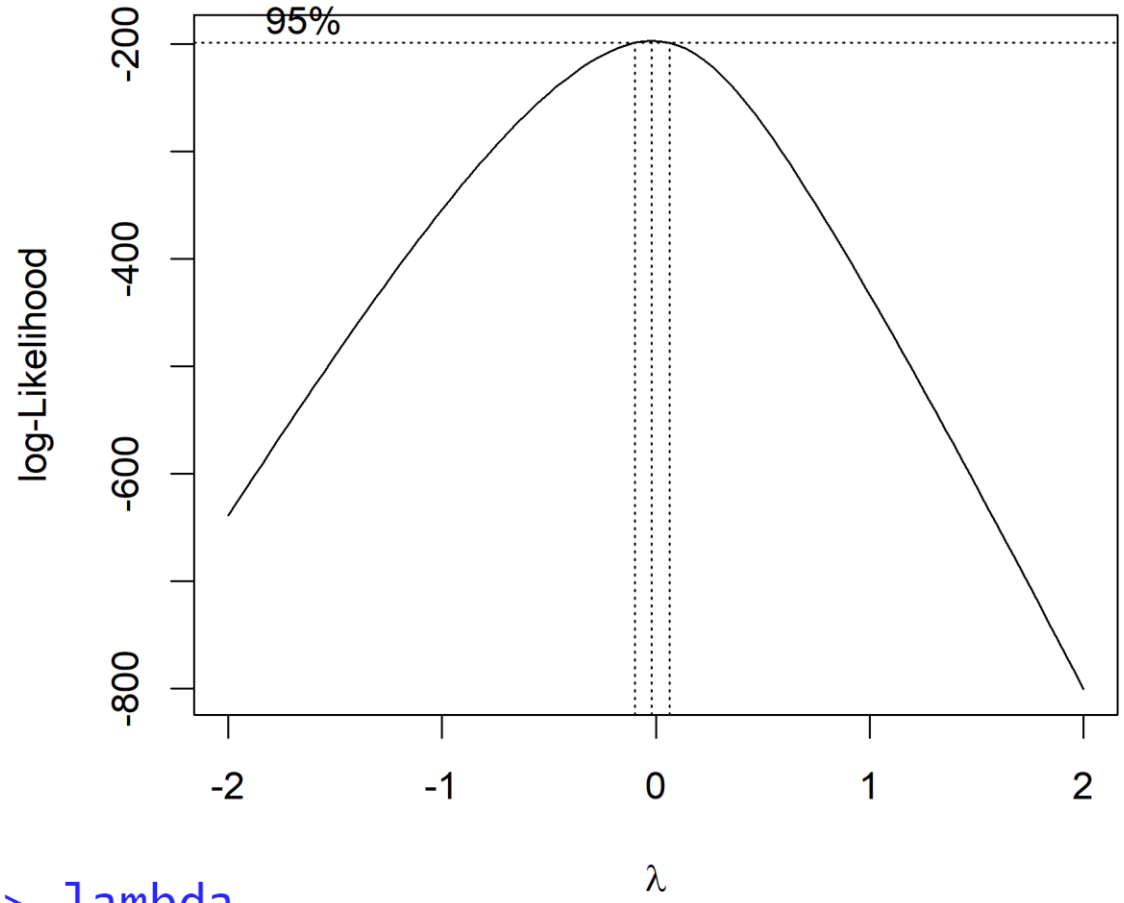






- Normality and constant variance seem problematic
- Remedy: Transformation of response?

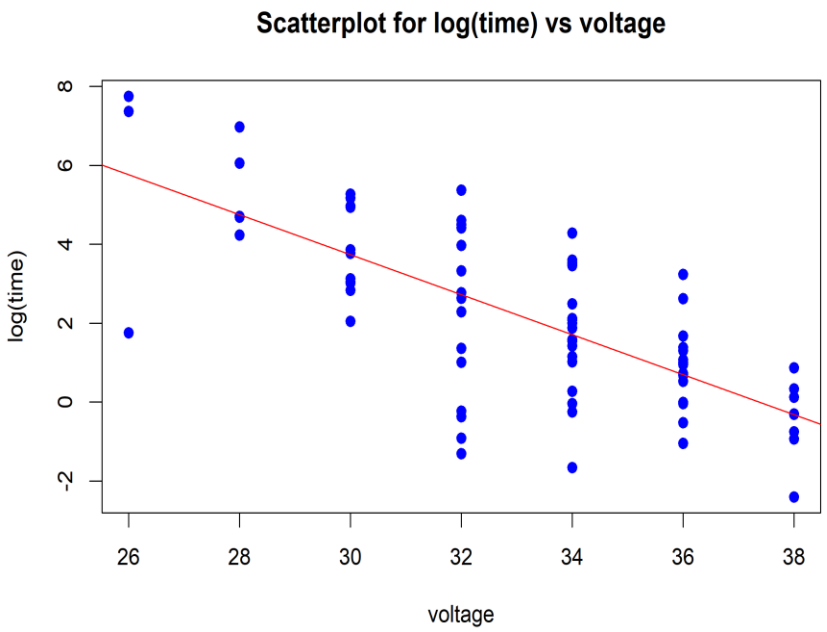
Box-Cox transformation

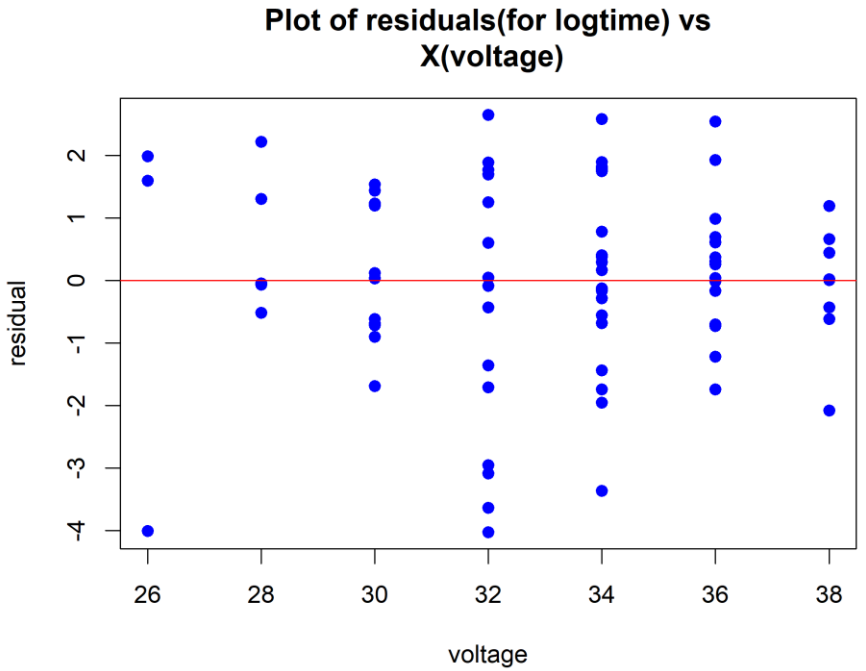


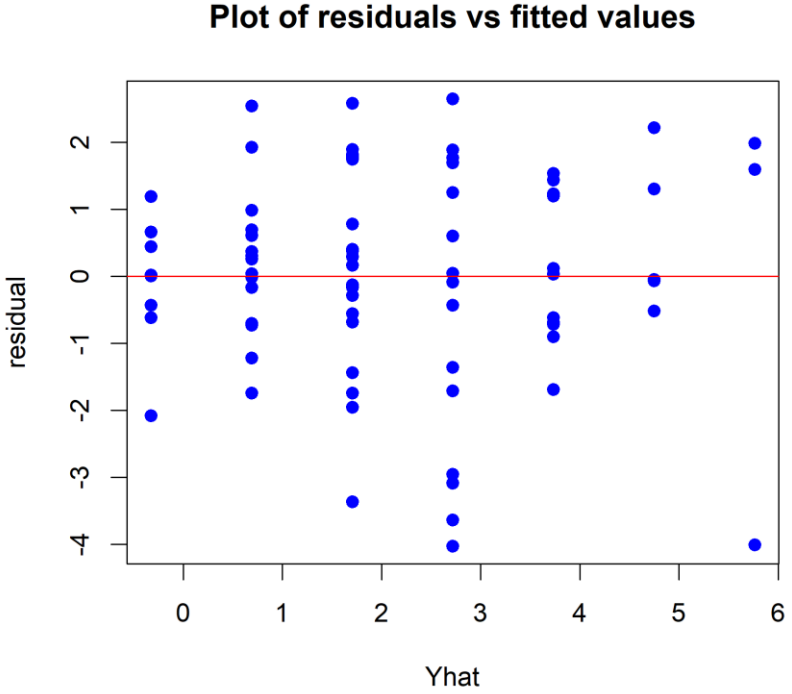

```
> lambda
[1] -0.02020202
```

- λ is close to 0 so I would try the ln transformation of Y, that is log(time)

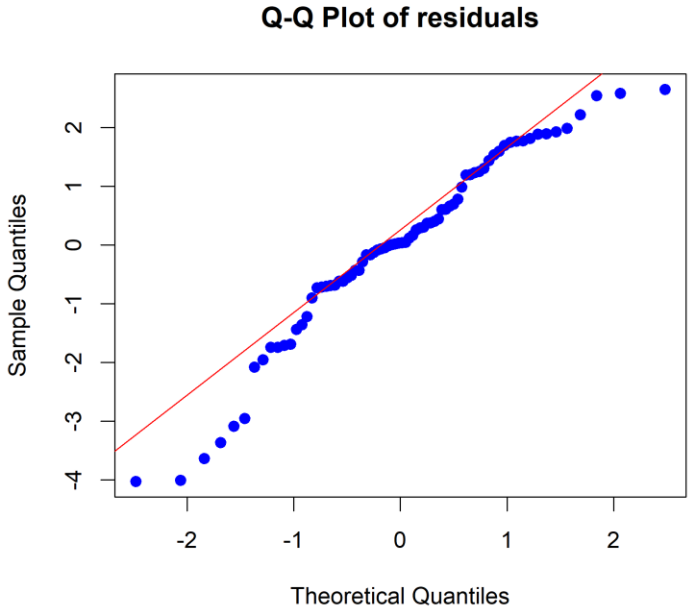
Transform and refit model







- Trend looks more linear
- Normality and constant variance seem to be improved even though normality could be better



Model interpretation

```
Call:
lm(formula = (log(time)) ~ voltage)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0291 -0.6919  0.0366  1.2094  2.6513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.9555     1.9100   9.924 3.05e-15 ***
voltage      -0.5074     0.0574  -8.840 3.34e-13 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.56 on 74 degrees of freedom
Multiple R-squared:  0.5136,    Adjusted R-squared:  0.507
F-statistic: 78.14 on 1 and 74 DF,  p-value: 3.34e-13
```

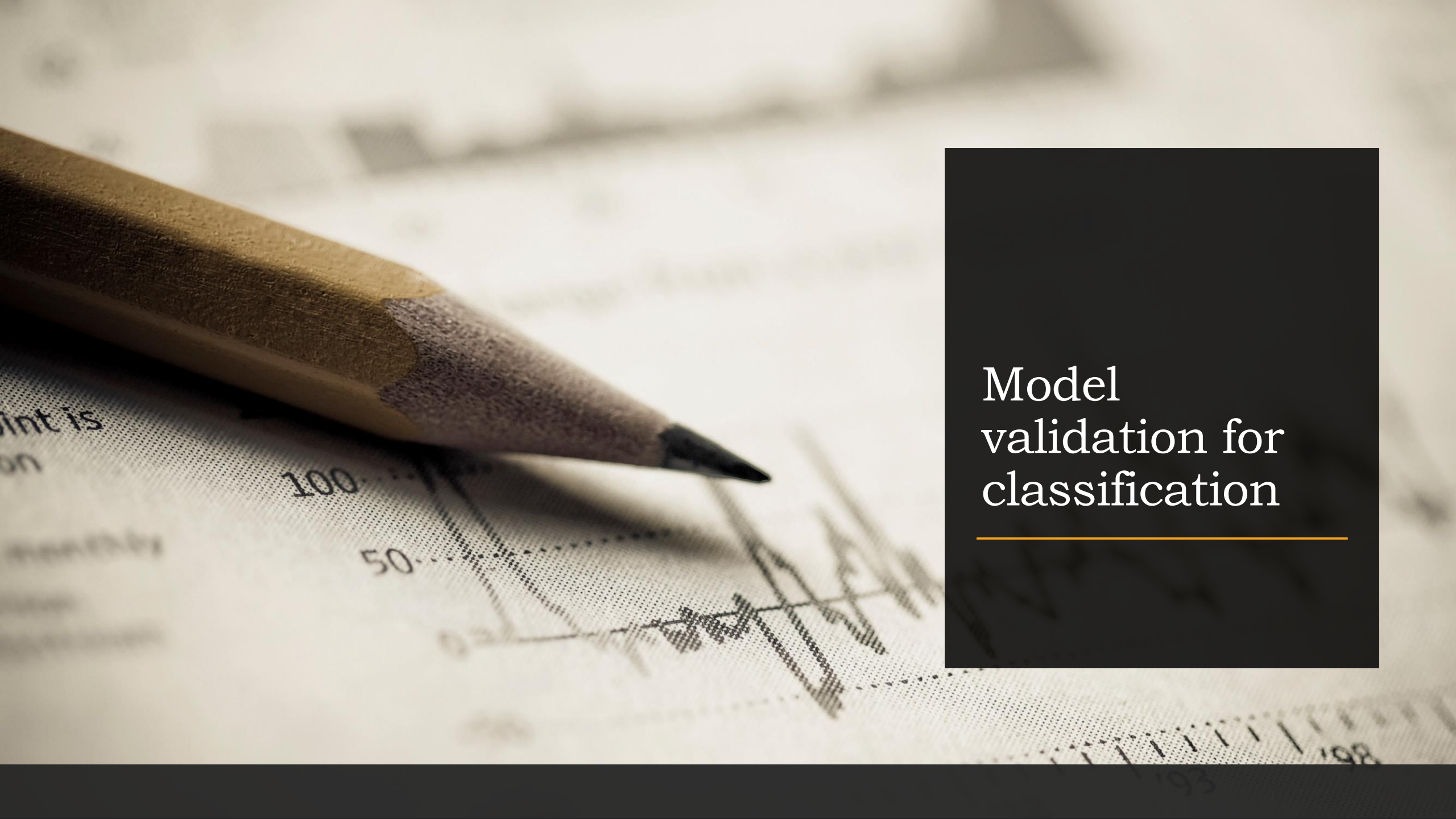
- There is a significant linear relationship between the (natural) log breakdown time and the voltage (p-value <0.001).
- New regression equation is $\widehat{\log(time)} = 18.95546 - 0.50736 * \text{voltage}$
- As voltage increases by one kV, we expect the log breakdown time to decrease by 0.507 minutes.

OR

- As voltage increases by one kV, we expect a multiplicative change of exp(-0.507)=0.60 in the breakdown time (i.e., a 40% decrease).

Summary of remedial measures

1. Non- linear relationship
   - Sometimes a transformation on X will fix this. Use non linear regression methods
2. Non- constant variance
   - Sometimes a transformation on Y will fix this. Non- normality
   - Sometimes a transformation on Y will fix this.
3. Non-constant variance and non-normality often go together.
4. Box-Cox Transformations – Help with transformations on Y
5. Sometimes a transformation on both X and Y may help.
6. Remember - Assumptions still need to be satisfied (on the transformed scale) if we are to use linear regression model.
7. So, we must always recheck diagnostic plots after transforming any variable.



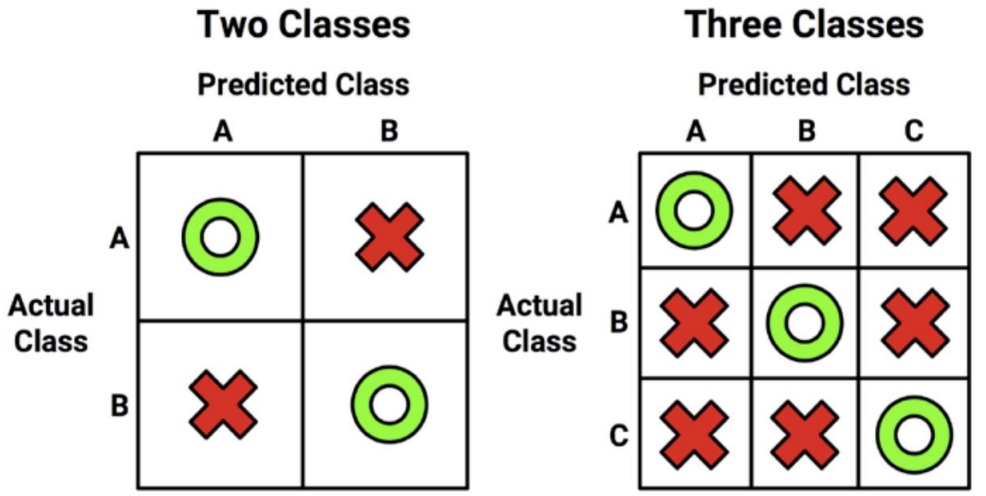
Model validation for classification

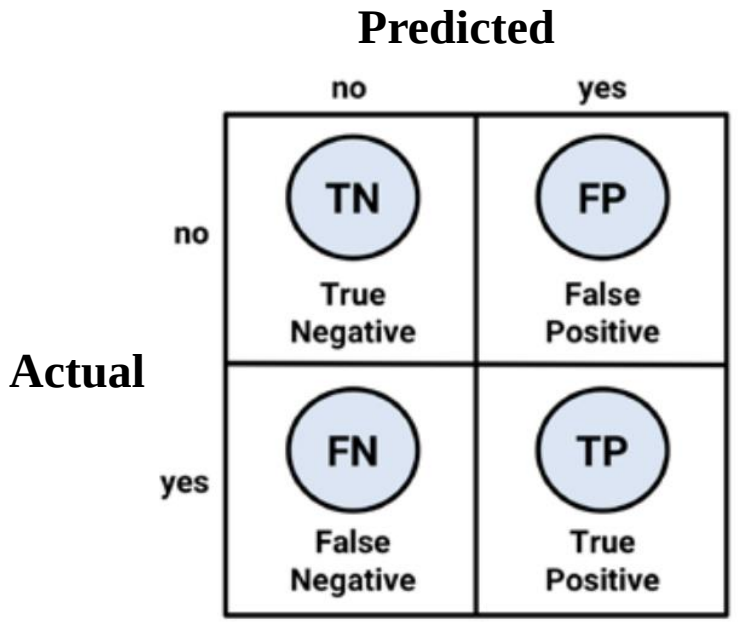
Cross validation

- Cross-validation in machine learning is a technique used to evaluate how well the model generalizes to new, unseen data and reduces the likelihood of overfitting.
- Cross-validation allows you to test the model's performance on different subsets of your data, providing a more robust evaluation of its predictive ability.
- In cross-validation, a dataset is split into several parts or "folds." The model is trained on all but one of these folds and then tested on the remaining fold. This process is repeated so that each fold serves as the test set exactly once. The model's performance metrics from each iteration are averaged to get a more reliable estimate of its effectiveness.
- Performing cross-validation at the model building stage is a good practice.

Confusion matrix

- A confusion matrix organizes predictions in relation to their actual values.
  - One dimension represents predicted value categories.
  - The other dimension represents actual value categories.
- The relationship between positive class and negative class predictions can be depicted as a 2x2 confusion matrix that tabulates whether predictions fall into one of four categories:
  - True positive (TP): Correctly classified as the class of interest
  - True negative (TN): Correctly classified as not the class of interest
  - False positive (FP): Incorrectly classified as the class of interest
  - False negative (FN): Incorrectly classified as not the class of interest





Other metrics

Accuracy

- Calculated as the proportion of the number of correct predictions (both true positives and true negatives) to the total number of predictions made.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- High accuracy indicates that the model is effective at classifying instances correctly

Error rate

- The proportion of incorrectly classified examples, that is

$$\text{error rate} = \frac{FP + FN}{TP + TN + FP + FN} = 1 - \text{accuracy}$$

- Low error rate indicates that the model is effective at classifying instances correctly

Other metrics

Sensitivity

- The sensitivity of a model (also called the true positive rate), measures the proportion of positive examples that were correctly classified, that is

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- High accuracy indicates that the model is effective at classifying instances correctly

Specificity

- The specificity of a model (also called the true negative rate), measures the proportion of negative examples that were correctly classified

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Other metrics

Precision

- The precision (also known as the positive predictive value) is defined as the proportion of positive predictions that are truly positive; in other words, when a model predicts the positive class, how often is it correct?
  - A precise model will only predict the positive class in cases very likely to be positive, making it trustworthy.
  - If the model was very imprecise, the results would be less likely to be trusted

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall

- Recall is also known as sensitivity or the true positive rate.
- It is a metric that measures the proportion of actual positives that are correctly identified by the model.
- It is calculated as the number of true positives divided by the sum of the true positives and the false negatives.
- High recall indicates that the model is good at capturing the positive cases.
- It is particularly important in situations where missing a positive case (like a disease in medical testing) would have serious consequences.

Other metrics

F- measure (F1 Score)

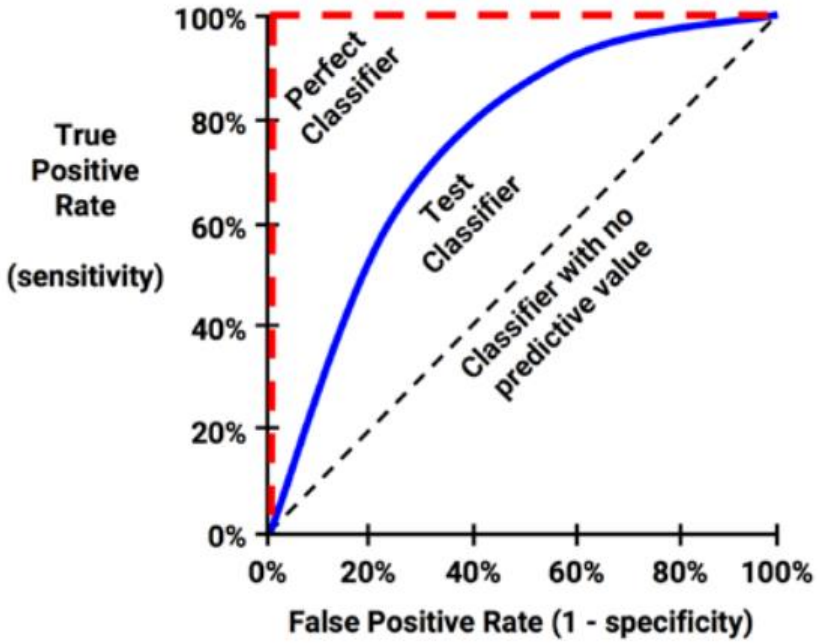
- The F1 Score is a statistical measure used to evaluate the accuracy of a classification model.
- It is the harmonic mean of precision and recall, providing a balance between these two metrics.
- This score is particularly useful when dealing with imbalanced datasets, as it accounts for both false positives and false negatives.
- High F1 Scores indicate a model with both high precision and high recall.

$$\text{F - measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{recall} + \text{precision}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

Other metrics

ROC Curve and AUC

- The ROC curve, or Receiver Operating Characteristic curve, is a graphical representation used to evaluate the performance of a classification model.
- It plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various threshold settings.
- The closer the curve is to the perfect classifier, the better it is at identifying positive values. This can be measured using a statistic known as the area under the ROC curve (AUC).
- The AUC treats the ROC diagram as a two-dimensional square and measures the total area under the ROC curve. AUC ranges from 0.5 (for a classifier with no predictive value), to 1.0 (for a perfect classifier)
- The area under the ROC curve (AUC) provides a measure of the model's ability to distinguish between classes.
- A larger AUC value indicates better model performance.
- The ROC curve is very useful for evaluating models on imbalanced datasets.



UP NEXT! Feature Selection and Regularization

Week 5

Feature Selection and Regularization in Regression

Ovefitting

Overfitting

- Overfitting occurs when a model learns both the underlying pattern and noise in the training data to such an extent that it performs poorly on unseen data.
- This makes the model highly accurate on its training data but less effective on new, unseen data, limiting its predictive power and generalizability.
- Consequences of overfitting include unreliable predictions and inferences.
- Signs of overfitting include:
  - High training accuracy but dropping validation/test accuracy
  - High variance across cross-validation folds
- Regularization and feature selection are essential in machine learning to address overfitting and improve model generalizability.
  - They aim to create simpler, more interpretable models that perform better on unseen data, ensuring that models are not just memorizing the training data but performing well with new data.

Feature Selection Techniques

- Feature selection involves identifying and using only the most relevant features, reducing model complexity, and improving performance.

- **Algorithm 6.1** *Best subset selection*

  1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
  2. For $k = 1, 2, \ldots p$:
     (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
     (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
  3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2.

- Pick model with smallest AIC, BIC, Mallow's C_p value less than but close to p+1 (number of variables in the model plus one), highest adjusted R^2.

Feature Selection Techniques

- **Automated feature selection:** Forward selection, Backward selection, Stepwise selection.
- **Another alternative** to the afore mentioned approaches: we can directly estimate the test error using the validation set and cross-validation methods
  - This procedure has an advantage relative to AIC, BIC, C_p , and adjusted R^2, in that it provides a direct estimate of the test error, and makes fewer assumptions about the true underlying model.

Introduction to Regularization

- An alternative to subset selection is Regularization.
- Here, we fit a model containing all p predictors using a technique that constrains the coefficient estimates, or equivalently, that shrinks the coefficients estimates towards zero.
  - Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero.
  - Hence, shrinkage methods can also perform variable selection.
- Also known as shrinkage methods.
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.
- The three best-known techniques for shrinking the regression coefficients towards zero are ridge regression, lasso and elastic net regression.

L2 Regularization (Ridge)

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- **Ridge regression** is very similar to least squares, except that the coefficients ridge regression are estimated by minimizing a slightly different quantity.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately.

- The shrinkage penalty is applied to $\beta_0, \beta_1, \ldots, \beta_p$, but not to the intercept β_0. We want to shrink the estimated association of each variable with the response; however, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when $x_{i1} = x_{i2} = \cdots x_{ip} = 0$

L2 Regularization (Ridge)

- Unlike least squares, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates, for each value of λ.
- Selecting a good value for λ is critical; and cross-validation is useful in finding the optimum value of λ.
- Ridge particularly useful for dealing with multicollinearity among predictors.
  - Multicollinearity occurs when predictor variables in a regression model are highly correlated.
  - This condition can lead to unreliable and unstable estimates of regression coefficients, making it difficult to determine the effect of individual predictors on the response variable.
  - High multicollinearity can also inflate the variance of the coefficient estimates, which may result in a failure to identify significant predictors.
  - Multicollinearity is a problem in your regression model if you find the Variance Inflaion Factor and it is greater than 10(some books say >5)

L2 Regularization (Ridge)

Con of Ridge regression

- Unlike best subset, forward, stepwise, and backward selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.
- The penalty $\lambda \sum \beta_j^2$ will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$) - will not result in exclusion of any of the variables.
- This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of variables p is quite large.
- It is best to apply ridge regression after scaling the predictors (dividing each predictor by its the standard deviation)

L1 Regularization (Lasso)

- The Least Absolute Shrinkage and Selection Operator (Lasso regression) is an alternative to the ridge regression that overcomes the problem of keeping all variables in the model.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

- Lasso can not only reduce overfitting but also perform feature selection by driving some coefficients to exactly zero, thus excluding them from the model altogether, hence identifies significant predictor variables.
- In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
- Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size, even though the number of predictors that is related to the response is never known beforehand.
- Cross-validation can be used in order to determine which approach is better on a particular data set.

Elastic Net Regression

- It is might sometimes be easy to choose between Lasso or Ridge based on knowledge of the variables in the dataset but when we have large number of parameters, how do we know which to choose in advance?
- Elastic Net regression is a combination of L1 and L2 regularization.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=1}^{p}\beta_j^2,$$

- We use cross validation on different sets of λ_1 and λ_2 to find the best λ values for the given data set.
- When $\lambda_1 = 0$ and $\lambda_2 = 0$, we have the OLS equation.
- When $\lambda_1 > 0$ and $\lambda_2 = 0$, we have lasso regression.
- When $\lambda_1 = 0$ and $\lambda_2 > 0$, we have ridge regression.
- When $\lambda_1 > 0$ and $\lambda_2 > 0$, we have elastic net regression.
- The method is useful when there is multicollinearity

Week 6

Generalized Linear Models

Introduction to Generalized Linear Models (GLMs)

- Generalized Linear Models (GLMs) extend traditional linear regression to accommodate response variables that do not necessarily follow a normal distribution.
- This flexibility allows GLMs to model a wide variety of data types and distributions, making them a powerful tool in both statistics and machine learning.
- Unlike linear regression models, which assume that the response variable is Normally distributed, GLMs allow for the response variable to follow different distributions such as Binomial, Poisson, and Exponential, among others.

Generalized Linear Models (GLMs)

- **Random Component:** This refers to the probability distribution of the response variable (Y).
  - In GLMs, the response variable can follow various distributions, each belonging to the exponential family of distributions, such as normal, binomial, Poisson and gamma.
  - The choice of distribution depends on the nature of the data and the research question.

- **Systematic Component:** This component is a linear combination of the explanatory variables $(X_1, X_2, \dots, X_n)$ and their associated coefficients $(\beta_1, \beta_2, \dots, \beta_n)$ similar to linear regression.
  - It is represented as $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$, where η is known as the linear predictor.
  - The systematic component models the effect that the predictors have on the response variable.

- **Link Function (g):** The link function provides the relationship between the linear predictor and the mean (E(Y)) of the response variable's distribution.
  - It transforms the expected value of the response variable (E(Y)) into the linear predictor (η).
  - Common link functions include the identity link for normal distributions, logit link for binomial distributions, and log link for Poisson distributions.

POISSON REGRESSION

Use Case: Modeling Count Data or Rates

- Poisson regression is for modeling count data or the rate at which events occur. This makes it particularly useful in fields such as public health (e.g., number of disease cases in a population), insurance (e.g., number of claims), and traffic management (e.g., number of accidents).
- It is applied when the data represent counts or the number of occurrences of an event within a fixed period of time or space, and these counts are assumed to follow a Poisson distribution.
- The counts should be non-negative integers, and the events should occur independently of one another.

Mathematics: Introduction to the Poisson Distribution and Its Properties

- The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, assuming these events occur with a known constant mean rate and independently of the time since the last event.
- The probability mass function (PMF) of the Poisson distribution for observing

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \text{ for y=0,1,2,…}$$

where λ is both the mean and the variance of the distribution, indicating the number of times an event is expected to occur during the given interval.

The parameter λ must be positive.

Interpretation: Explaining the Interpretation of Coefficients as Log-Counts

- The model is typically expressed in the form of

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon,$$

Where λ is the expected count of the dependent variable, and $X_1, X_2, \ldots, X_n$ are the independent variables.

- In Poisson regression, the coefficients represent the change in the expected log count of the dependent variable for a one-unit change in the predictor, holding all other predictors constant.

To interpret the coefficients (β) of the Poisson regression model:

- A one-unit increase in X_i is associated with a change in the log count of the response variable by β_i, holding all other predictors constant.
- e^{β_i} gives the factor change in the count. For example, if $\beta_i = 0.2$, then a one-unit increase in X_i multiplies the count(response) by $e^{0.2} \approx 1.22$, indicating an approximate 22% increase in the count.
- This interpretation allows us to understand the impact of predictors on the rate of occurrence of the events being modeled.

Interpretation: Explaining the Interpretation of Coefficients as Log-Counts

- Suppose there is a categorical variable (e.g. Regions with categories North, South, East, West) in the Poisson regression model, with the Northern region as baseline.

- Supposing $X_1 = 1$ if Region is South and $X_1 = 0$, otherwise. Then
  - $\beta_1 > 0$ means the expected log count of the response variable for the South region is β_1 higher than for the North region, holding all other predictors constant.
  - $\beta_1 < 0$ means the expected log count of the response variable for the South region is β_1 lower than for the North region, holding all other predictors constant.

Interpretation: Explaining the Interpretation of Coefficients as multiplicative effect

- e^{β_1} represents the factor by which the expected count for the South region is higher or lower than the expected count for the North region.
  - $e^{\beta_1} > 0$ means the South region has an expected count of the response variable that is e^{β_1} times the count for the North Region, holding all other predictors constant. For example, if e^{β_1} =1.5, it means the expected count in the South is 50% higher than in the North.
  - $e^{\beta_1} < 0$ means the South region has an expected count of the response variable that is e^{β_1} times the count for the North Region, holding all other predictors constant. For example, if e^{β_1} =0.8, it means the expected count in the South is 20% lower than in the North.
  - $e^{\beta_1} = 1$ means there is no difference in the expected count between the South and North regions.

Diagnostics: Checking for Model Assumptions and Residual Analysis for poisson regression

- Checking for Overdispersion: Overdispersion occurs when the observed variance of the count data is greater than the mean, which can violate Poisson regression assumptions.
- Ways to check for overdispersion in the Poisson regression model are outlined in R (Lines 25-54):
- Residual Analysis: Plot the residuals against fitted values and each predictor. Look for patterns or trends that might suggest non-linearity or other model specification issues. A random scatter of points is a good sign.

Steps for Manual K-Fold Cross-Validation:

1. Split Your Data: Divide your data into k folds. For example, if you choose 5-fold cross-validation, your data will be divided into 5 parts.
2. Loop Through Folds: For each fold, use it once as a testing set while the remaining k-1 folds form the training set.
3. Fit the Model on Training Data: For each split, fit a negative binomial regression model to the training data.
4. Predict on Testing Data: Use the fitted model to make predictions on the testing data.
5. Calculate Performance Metrics: After making predictions, calculate performance metrics (e.g., RMSE, mean absolute error, MSE) for each fold.
6. Aggregate Results: Average the performance across all folds to get a general estimate of the model's performance.

Other Generalized Linear models

- Linear Regression: Y has a Normal distribution with identity link.
- Logistic Regression: Y has a Binomial distribution/ Multinomial distribution with logit link, used for binary outcomes.
- Poisson Regression: Y has a Poisson distribution with log link, used for count data.
- Negative Binomial Regression is particularly suited for count data where the variance exceeds the mean, a condition known as overdispersion.
  - The link function most commonly used is the logarithm,

$$\log[E(Y|X)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

- Gamma regression is suitable for modeling positive continuous data, especially when the data are skewed and not normally distributed.
  - The link function most commonly used in gamma regression is the logarithm,

$$\log[E(Y|X)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$