Week 5:
Linear Regression

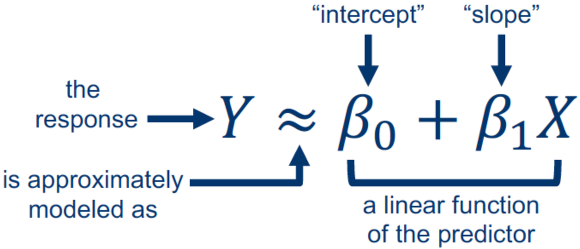
CMPS 320: Machine Learning

Outline

1. Simple Linear Regression
2. Multiple Linear Regression
3. Other Considerations in the Regression Model
   - Qualitative Predictors
   - Extending the linear model
   - Potential problems in linear regression

Simple Linear Regression

- An approach for predicting a quantitative response Y on the basis of a single predictor variable X.
- Assumption: there is a linear relationship between X (the predictor) and Y (the response)
- Mathematically, we can write this linear relationship as:



- β_0 and β_1 which are unknown constants are referred to as the model constants or parameters

Simple Linear Regression

- Given set of data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- The goal is to find estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, \dots, n$$

  the linear model fits the data well.

Residuals and Residual Sum of Squares (RSS)

- Let $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X.
- The ith residual is defined as

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

  i.e. the difference between observed and predicted response.
- The residual sum of squares (RSS) as defined as

$$\begin{aligned} RSS =& e_1^2 + e_2^2 + \ldots + e_n^2 \\ RSS =& (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + \\ & (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \end{aligned}$$

Minimizing RSS: least squares

- Using least square the goal is to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes RSS.
- Using calculus the minimizers are:

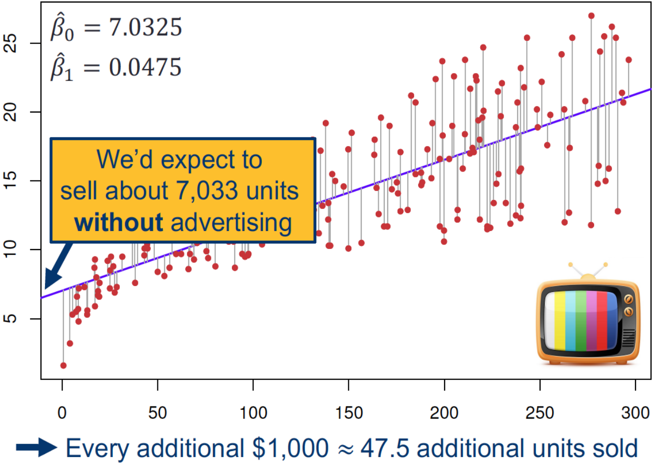
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

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

Example–Advertising

- For the Advertising data, the least squares fit for the regression of sales onto TV is shown.



Accuracy of the Coefficient Estimates–standard error

- The true relationship between X and Y takes the form, $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.
- If f is to be approximated by a linear function, then we can write this relationship as

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- When estimating the population mean μ of a random variable Y, natural question is as follows:
  - how accurate is the sample mean $\hat{\mu}$ as an estimate of μ?
- We answer this by computing the standard error of $\hat{\mu}$ as:

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

  where σ is the standard deviation of the population and n is the number of samples.
- Note: the error gets smaller as the sample size increases.

Accuracy of the Coefficient Estimates–standard error

- In a similar vein, we can wonder how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β_0 and β_1 .
- The standard errors associated with $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed using the formulas:

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

,

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

where $\sigma^2 = Var(\epsilon)$ and ϵ is the error.
- In the formula above, $\hat{\beta}_1$ is smaller when the x_i are more spread out
- Similarly, $\hat{\beta}_0$ would be the same as $SE(\hat{\mu})$ if $\bar{x}$ were zero (in which case $\hat{\beta}_0$ would be equal to $\bar{y}$).

Accuracy of the Coefficient Estimates–residual standard error

- In general, σ^2 is not known, however it can be estimated from the data.
- The estimate of σ is known as the **residual standard error** (RSE), and is given by the formula:

$$RSE = \sqrt{\frac{RSS}{(n-2)}}$$

- 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
- For linear regression, the 95% confidence interval for β_1 and β_0 approximately takes the form

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1) \text{ and } \hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0)$$

Using SE for hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the hypothesis coefficients.
- Hypothesis test involves testing the null hypothesis of

$$H_0 : \text{ There is no relationship between } X \text{ and } Y$$

versus the alternative hypothesis

$$H_a : \text{ There is some relationship between } X \text{ and } Y$$

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0$$

Example–Advertising

- The Table provides details of the least squares model for the regression of number of units sold on TV advertising budget 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

- The coefficients for $\hat{\beta}_0$ and $\hat{\beta}_1$ are very large relative to their standard errors, so the t-statistics are also large.
- At 5% significance level (i.e. $\alpha = 0.05$), we reject the null hypothesis – that is, we declare a relationship exist between TV advertising and Sales – since the p-value is less than 0.05.

Assessing model accuracy– RSE

- Once we have rejected the null hypothesis in favor of the alternative hypothesis, it is natural to want to quantify the extent to which the model fits the data.
- The quality of a linear regression fit is typically assessed using the R^2 **statistic**.
- R^2 statistic measures the proportion of variance explained by the model.
- R^2 takes on a value between 0 and 1, and is independent of the scale of Y.
- R^2 statistic close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.

Assessing model accuracy– R^2

- A number near 0 indicates that the regression did not explain much of the variability in the response;
  - this might occur because the linear model is wrong, or the inherent error σ^2 is high, or both.

R^2 is one such measure:

variance not explained after regression

$$R^2 = 1 - \frac{RSS}{\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{total variance in the response}}}$$

Multiple Linear Regression

- The multiple linear regression model takes the form

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

  where X_j represents the jth predictor and β_j quantifies the association between that variable and the response.
- The coefficient β_j is interpreted as the average effect on Y of a one unit increase in X_j , holding all other predictors fixed.
- The regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are unknown and estimated by using least squares (same as in simple linear regression)
- The coefficient β_j is interpreted as the average effect on Y of a one unit increase in X_j , holding all other predictors fixed.

Questions we ask in Multiple Linear Regression

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- Do all the predictors help to explain Y , or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Qualitative Predictors

- In our discussion so far, we have assumed that all variables in our linear regression model are quantitative.
- However in practice, this is not necessarily the case; often some predictors are qualitative.
- Examples of qualitative variables are sex, marital status, political affiliation etc

Two-level Qualitative Predictors

- Suppose that we wish to investigate differences in credit card balance between males and females, ignoring the other variables for the moment.
- We create an indicator or dummy variable that takes on two possible numerical values.
- Based on the gender variable, we can create a new variable that takes the form:

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

- This results in the regression model:

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

A note on dummy variables

- The decision to code females as 1 and males students as 0 is arbitrary.
  - It has no effect on model fit, or on the predicted values.
- Alternatively, instead of a 1/0 coding scheme, we could create a dummy variable:

$$x_i = \begin{cases} 1 & \text{if } ith \text{ person is female} \\ -1 & \text{if } ith \text{ person is male} \end{cases}$$

- This results in the regression model:

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

- Using this coding scheme, the final predictions for the credit balances of males and females will be identical to the previous scheme.
- The only difference is in the way that the coefficients are interpreted.

Extending the linear model

- The linear regression model provides nice, interpretable results and is a good starting point for many applications.
- We outline some classical approaches for extending the linear model:
- Linear Relationships: Allowing for interaction effects

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

- Nonlinear Relationships: Using polynomial regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

- This is still a linear model since we can rewrite:

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

where $X_2 = X_1^2$

Non-linearity of the Data

- The linear regression model assumes that there is a straight-line relationship between the predictors and the response.
- If the true relationship is far from linear, then all of the conclusions that we draw from the fit are suspect.
- The prediction accuracy of the model can be significantly reduced.
- Residual plots are a useful graphical tool for identifying non-linearity.
- If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, such as $\log X$, $\sqrt{X}$, and X^2, in the regression model.

Correlation of Error Terms

- The linear regression model assumes that the error terms are uncorrelated.
- If these terms are correlated, the estimated standard error will tend to underestimate the true standard error.
- As a result, confidence and prediction intervals will be narrower than they should be.
- **Question**: Why might correlations among the error terms occur?
- Such correlations frequently occur in the context of time series data, which consists of observations for which measurements are obtained at discrete points in time.
- In the time-sampled case, we can plot the residuals from our model as a function of time.
- Uncorrelated errors = no discernable pattern

Non-constant variance of error terms

- The linear regression model assumes that the error terms have constant variance:

$$Var(\epsilon_i) = \sigma^2$$

- Often not the case (e.g. error terms might increase with the value of the response)
- Non-constant variance in errors = heteroscedasticity
- How to identify heteroscedasticity.
  - The residuals plot will show a funnel shape
- How to fix heteroscedasticity.
  - transform the response using a concave function (like log or sqrt)
  - weight the observations proportional to the inverse variance

Outliers

- **Outlier**: an observation whose true response is really far from the one predicted by the model
- Sometimes indicate a problem with the model (i.e. a missing predictor), or might just be a data collection error.
- Can mess with R^2, which can lead us to misinterpret the model's fit.
- How to identify outliers?
  - Residual plots can help identify outliers, but sometimes it's hard to pick a cutoff point (how far is "too far"?)
- How to fix outliers?
  - Divide each residual by dividing by its estimated standard error (studentized residuals), and flag anything larger than 3 in absolute value.

High leverage points

- Outliers = unusual values in the **response**
- High leverage points = unusual values in the **predictor(s)**
- The more predictors you have, the harder they can be to spot (why?)
- These points can have a major impact on the least squares line (why?), which could invalidate the entire fit
- How to identify high leverage points
  - Compute the leverage statistic

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

  - A large value of this statistic indicates an observation with high leverage.

Collinearity

- Problems can also arise when two or more predictor variables are closely related (correlated) to one another
- The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.
- How to detect collinearity:
  - Look at the correlation matrix of the predictors
  - An element of this matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data.
- It is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation. We call this situation **multicollinearity**.

Collinearity

- A better way to assess **multicollinearity** is to compute the variance inflation factor (VIF).
    - VIF quantifies how much the variance is inflated.
- The smallest possible value for VIF is 1, which indicates the complete absence of collinearity.
- As a rule of thumb, a VIF value that exceeds 5 indicates a problematic amount of collinearity.
- Dealing with collinearity:
    - Drop one of the problematic variables from the model.
        - ★ The decision of which one to remove is often a scientific or practical one.
    - Combine the collinear variables together into a single predictor