STA 4241 Lecture, Week 2

August 30th, 2021

Overview of what we will cover

- Review of linear regression
  - Simple linear regression
  - Multiple linear regression
  - Estimating coefficients
  - Hypothesis testing
  - Removing linearity and additivity assumptions

Linear regression

- Suppose again that we are interested in the following model:

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

- Linear regression broadly refers to methods that assume $f(\cdot)$ to be linear in the predictor X
- It is important to have a strong understanding of linear regression before discussing more complex methods in this course

Why linear regression is important

- Many of the more complex methods we will see in this course are extensions of, or are rooted in linear regression
- We can actually create quite flexible models just within the scope of linear regression
- Additionally, the linear model is frequently a good approximation to the true regression function
    - Don't always need the fancier models

Why linear regression is important

- Linear regression is also extremely easy to implement
- Very interpretable results
  - Coefficients in the model have a nice interpretation
  - Inference on coefficients is straightforward
  - Easy to tell which predictors are important for predicting the outcome
- Widely studied and very well understood approach

Simple linear regression

- Let's first discuss linear regression with only one predictor, X
  - Called simple linear regression
- The model is therefore

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

- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and is assumed independent of X
  - We assume normality throughout, but it is not necessary
- σ^2 is commonly called the residual variance of the model

Simple linear regression

- The book sometimes refers to this model as

$$Y \approx \beta_0 + \beta_1 X$$

- It is more precise to use the equation on the previous slide or to say

$$E[Y|X] = \beta_0 + \beta_1 X$$

- And the residual variance is defined as

$$\text{Var}[Y|X] = \sigma^2$$

Interpretation of parameters

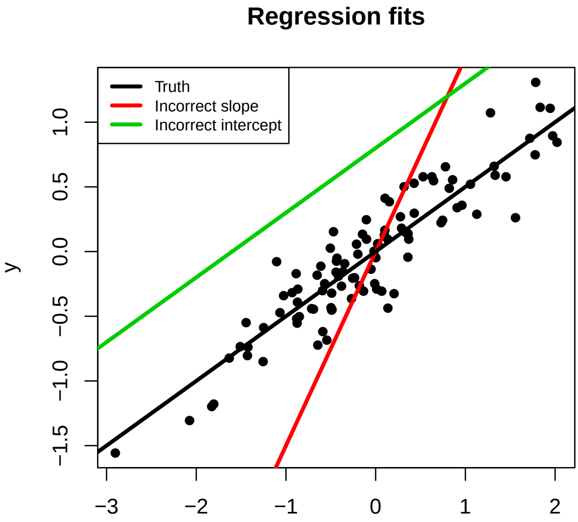
- β_0 is the expected value of the outcome when $X = 0$
  - Only interpretable when X can reasonably take the value 0
  - For instance, suppose X is BMI, which can never be 0
  - We can always center X to make the intercept interpretable
- β_1 is the expected change in the outcome for a one unit change in the predictor X

Simple linear regression

- Once we posit a model, we simply need to estimate the unknown parameters
- We want to find estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that fit the data well
- We will use the data given by (X_i, Y_i) for $i = 1, \ldots, n$
- Before going into mathematical details, let's think intuitively what we want the parameter values to be

Simple linear regression

- We want values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that $Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$ is small



Simple linear regression

- Define $e_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$ to be the ith residual
- We know we want e_i to be small, but how do we quantify this?
- The least squares criterion is the most common approach
- We want to find β_0 and β_1 that minimize

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} e_i^2 \equiv \text{RSS}$$

- The least squares estimator is the value $(\widehat{\beta}_0, \widehat{\beta}_1)$ that minimizes RSS

Simple linear regression

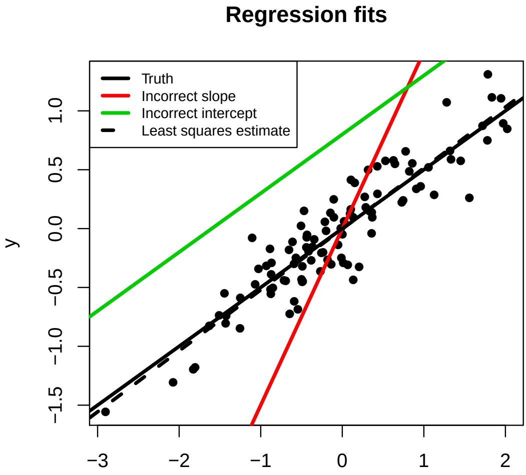
- It turns out the least squares solution has a really simple form

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

- Where $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$
- Least squares isn't the only criterion we could use to find the parameter estimates
- Could alternatively minimize the sum of the absolute residuals
    - Focus on least squares for now

Simple linear regression

- The least squares solution on this example is extremely close to the truth



Simple linear regression

- The true line is called the population regression line
- The difference between the population regression line and the least squares line is due to sampling variability
- The least squares line is an estimate of the true, unknown population line based on a sample of size n
- It can be shown that $E(\widehat{\beta}_0) = \beta_0$ and $E(\widehat{\beta}_1) = \beta_1$
    - Unbiased estimator

- Another quantity of interest is how close we expect our estimates to be on average from the truth
- Quantify this with the variance of these estimates

$$\mathrm{Var}(\widehat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right]$$

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

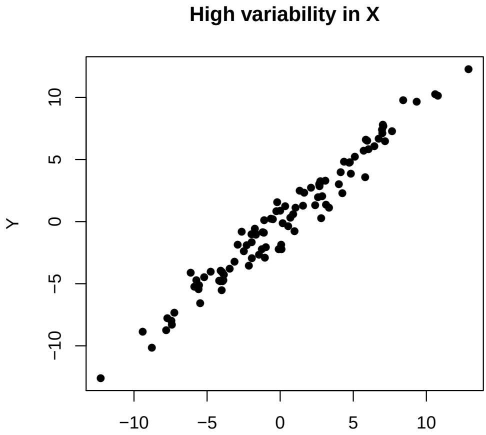
- And the standard errors are simply the square roots of these quantities

Simple linear regression

- These standard errors provide some intuition about the estimators
- Both estimators are more efficient (smaller standard errors) when there is more variability in X
  - $\sum_{i=1}^{n}(X_i - \overline{X})^2$ is large
- We can see visually why this is the case

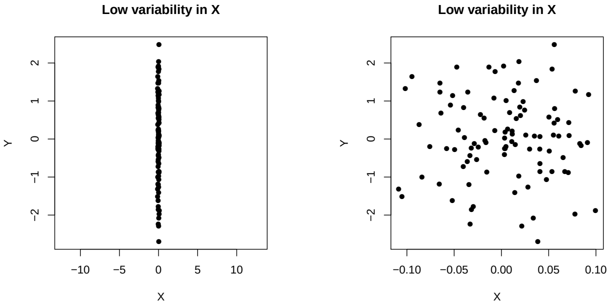
Simple linear regression

- Generate data from a linear regression model with lots of spread in the X variable
- Very easy to see the true line visually



Simple linear regression

- I generate data from the same regression but I only observe X values in a small range
- Much more difficult to estimate the unknown line



Simple linear regression

- In practice we don't know these standard errors
  - Residual variance σ^2 is not known
- We can estimate the standard errors by plugging in an estimate of σ^2

$$\widehat{\sigma}^2 = \frac{\mathsf{RSS}}{n-2}$$

- Denote these standard error estimates by $\widehat{\mathsf{SE}}(\widehat{\beta}_0)$ and $\widehat{\mathsf{SE}}(\widehat{\beta}_1)$

Simple linear regression

- Once we have estimates and standard errors for the parameters, we can construct confidence intervals and do hypothesis testing
- A $100(1-\alpha)\%$ confidence interval for β_1 can be constructed as

$$\widehat{\beta}_1 \pm t_{1-\alpha/2,n-2}\widehat{\mathsf{SE}}(\widehat{\beta}_1)$$

- where $t_{1-\alpha/2,n-2}$ is the $1-\alpha/2$ quantile of the t-distribution with $n-2$ degrees of freedom
- When n is large (above 30) this is well approximated by a normal distribution
- Same can be done for β_0

- Standard errors also allow us to perform hypothesis tests
- We are typically interested in whether there is any relationship between X and Y
- In our model, this is represented by the following null and alternative hypotheses

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0$$

- Don't typically perform hypothesis tests for the intercept

Simple linear regression

- Our estimate $\widehat{\beta}_1$ will never be exactly zero
- The standard error tells us if the difference is sufficiently far from zero to reject the null hypothesis
- Specifically we use the following statistic

$$t = \frac{\widehat{\beta}_1}{\widehat{\mathrm{SE}}(\widehat{\beta}_1)}$$

- Measures the number of standard deviations from zero

Simple linear regression

- Our goal with testing is to control the type I error at α
  - Probability that we reject H_0 under the null is α
- Under H_0 the statistic follows a t-distribution with $n-2$ degrees of freedom
- The p-value is the probability, under H_0, of observing a value as large or larger in absolute value than $|t|$

- We reject H_0 if the p-value is less than α
- Smaller p-values indicate more evidence against the null hypothesis
- While it is important to understand why these hypothesis tests work and how to perform them, R will output all relevant quantities such as the p-value and test statistic for you

Model accuracy

- We want to have some measure of how good our model is
  - How well does it fit the observed data
  - How well does it predict new data points
- We will look at two measures of model fit
  - RSE
  - R^2
- These are measures of how well the model fits our observed data
  - Does not measure how well it predicts new data points

Model accuracy

- The RSE is defined as

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$$

  where $\widehat{Y}_i$ is the predicted value from our model

- The RSE is an estimate of the residual standard error in our model
- Smaller values of RSE indicate the predicted values are closer to the truth, and our model fits the data well

Model accuracy

- What is considered a good or low value of RSE depends heavily on the data set and the scale of Y
- R^2 is a measure that always falls between 0 and 1, and is independent of the scale of Y
- The formula for R^2 is

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

where TSS is defined as

$$\text{TSS} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

Model accuracy

- TSS is a measure of how good our predictions would be if we did not include X
- RSS is necessarily less than TSS, therefore $R^2 > 0$
- If RSS is very low compared to TSS, that indicates that X greatly improves the predictions in the model
- For simple linear regression, $R^2 = \text{Cor}(X, Y)^2$

Model accuracy

- We must be careful with both of these measures
- These measure how well the model fits the training/observed data
- Does not measure predictive performance on testing/new data
- These measures are susceptible to overfitting
  - Not typically a problem for simple linear regression
  - Becomes a problem with nonlinear terms or many covariates

Multiple linear regression

- Frequently we don't have just one covariate X
- Now suppose we observe $[X_1, X_2, \ldots, X_p]$
- We are interested in fitting a model of the form:

$$E(Y|X_1, \ldots, X_p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- Sometimes it will be useful to use matrix notation

$$E(Y|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$$

where $\boldsymbol{X} = [1, X_1, \ldots, X_p]$

Multiple linear regression

- Why use multiple linear regression and not many simple linear regressions?
- For each covariate, we could fit

$$E(Y|X_j) = \beta_0 + \beta_1 X_j$$

- There are two major problems with this approach
  - How do I use the output from these models to predict Y for a given set of covariates?
  - If the covariates are correlated, individual models can be very misleading

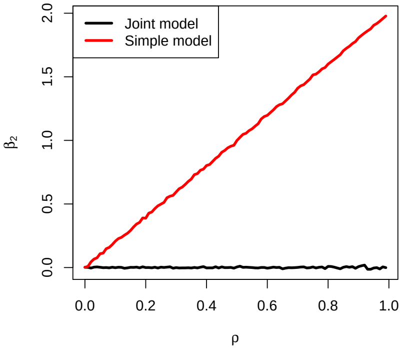
Multiple linear regression

- Suppose we have a simple example with two covariates X_1 and X_2
  - Correlation between X_1 and X_2 is ρ
- The true model is

$$E(Y|X_1, X_2) = 5 + 2X_1 + 0X_2$$

- We will fit two models
  - One that includes both covariates
  - One that only includes X_2
- Compare coefficients for the effect of X_2 under different ρ values

- The joint model correctly estimates a value very close to zero
- The simple model shows a strong association between X_2 and Y



Multiple linear regression

- This example highlights a key difference between marginal correlation and conditional correlation
- X_2 and Y are in fact correlated, marginally
- Correlation is only through X_1
- Once we condition on X_1, the correlation between X_2 and Y disappears
- The two models inherently answer different questions
  - Usually, the joint model is of more interest

Interpretation of parameters

- The parameters in the multiple linear regression parameters have a nice interpretation
- β_1 can be interpreted as the expected change in the outcome for a one unit change in X_1 if we fix the values of the remaining parameters
  - Conditioning on values of the other covariates
- The intercept is the average value of the outcome if we set all covariates to zero
  - Again only interpretable if zero is a reasonable value for the covariates

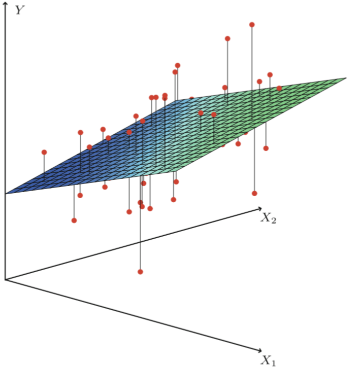
Estimating coefficients

- Now that we have a model, we need to estimate $\boldsymbol{\beta}$, the regression coefficients
- We again take the least squares approach
- We will aim to minimize

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i\beta)^2 = (\boldsymbol{Y} - \boldsymbol{X}\beta)^T(\boldsymbol{Y} - \boldsymbol{X}\beta)$$

Estimating coefficients

- Suppose we are interested in an example with only two covariates
- The regression model is now represented by a plane instead of a line
- Minimize squared difference between points and the plane



James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.

Estimating coefficients

- Let's use calculus to show that the least squares estimate is $\widehat{\beta} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{T}\boldsymbol{Y}$

Estimating coefficients

- We can also show that the variance of this estimator is given by $\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

Hypothesis testing

- One nice feature of linear regression is that there are many different hypotheses we can easily test
- Suppose we are interested if there is any relationship between the predictors and the outcome
- This corresponds to

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus

$$H_a : \text{At least one } \beta_j \text{ is nonzero}$$

Hypothesis testing

- The statistic for this test is given by

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

  where TSS $= \sum_{i=1}^{n}(Y_i - \overline{Y})^2$
  and RSS $= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$

- Note that for multiple linear regression, $\widehat{Y}_i = \boldsymbol{X}_i\widehat{\beta}$

Hypothesis testing

- The denominator satisfies

$$E[RSS/(n-p-1)] = \sigma^2$$

- Under H_0, the numerator satisfies

$$E[(TSS - RSS)/p] = \sigma^2$$

- This means that under the null hypothesis, we expect this statistic to be close to 1

- Under the alternative, we expect RSS to get smaller and therefore

$$E[(TSS - RSS)/p] > \sigma^2$$

- Large values of the F-statistic therefore provide support against the null hypothesis
- How large the F-statistic needs to be depends on the sample size, but R will output a p-value for you
  - F-statistic follows an F-distribution under H_0

- What if we only want to test whether a subset of the parameters are zero?
- Suppose we are interested in testing

$$H_0 : \beta_{j_1} = \cdots = \beta_{j_q} = 0$$

- And the alternative hypothesis is that one of these is nonzero
- We can easily change the F-statistic to account for this

Hypothesis testing

- We simply need to let RSS_0 be the residual sums of squares from the model that includes all covariates *except* the q covariates of interest
- The statistic then becomes

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

- We simply changed the null model from being the one with zero covariates to the one that included all covariates except the q of interest
- This statistic will be large if these additional q covariates greatly reduce the residual sums of squares

Hypothesis testing

- Why do we need the F-statistic? Can't we just use the p-values from each individual covariate to determine if any are important?

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1391272  0.0764830  -1.819   0.0708 .
x1          -0.0046568  0.0713657  -0.065   0.9481
x2          -0.0288001  0.0807171  -0.357   0.7217
...
x39         -0.0006744  0.0788347  -0.009   0.9932
x40          0.0466208  0.0740701   0.629   0.5300
```

- If any of them are significant, doesn't that imply that we can reject the null hypothesis that all are zero?

Hypothesis testing

- This approach will lead to false discoveries and poor type I error rates
- To see this, I simulated 1000 data sets with $p = 40$ covariates
- In all data sets, there is no relationship between $\boldsymbol{X}$ and Y
  - True values are $\beta_1 = \cdots = \beta_p = 0$ and H_0 is true
- Below is the type I error rate if I use 1) the F-statistic and 2) reject if any of the individual p-values are less than α

	F-test	Individual tests
type-I error	0.05	0.86

Hypothesis testing

- The more variables you include, the higher the chance of type-I error when you base the test on the individual p-values
- The F-test accounts for the number of variables and is unaffected
- The individual tests can be fixed by adjusting the cutoff value for the p-value that let's us deem a parameter significant
- If $p > n$, neither approach works, though we'll discuss that later in class

Model fit

- As with simple linear regression, we are interested in how well our model fits the data
- Both R^2 and RSE are applicable to both simple and multiple linear regression assuming minor tweaks to their formulae
- In simple linear regression, $R^2 = \text{Cor}(X, Y)^2$
- Now, we have that $R^2 = \text{Cor}(Y, \widehat{Y})^2$
- It still has the interpretation of being the percent of variability in Y that is explained by $\boldsymbol{X}$

Model fit

- The RSE is defined as

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$$

- The RSE is an estimate of the residual standard deviation in the model, σ

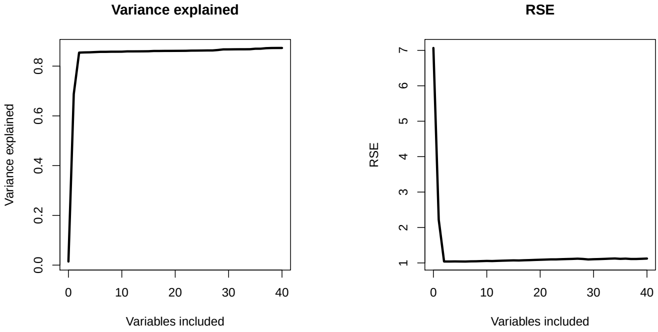
- Smaller RSE values are caused by predictions being closer to the true values, meaning that our model fits the observed data well

Model fit

- As we include more predictors into the model R^2 necessarily goes up
  - Even if these predictors are irrelevant
- RSS also necessarily goes down as we include more parameters
- RSE can either go up or down, depending on how much predictors change RSS
- We need to be aware of overfitting when using these as measures of model quality
  - Better to use out of sample or testing data to evaluate model performance

Model fit

- Suppose true model is $Y = 2X_1 + X_2 + \epsilon$, and $\text{Var}(\epsilon) = 1$
- Below is the R^2 and RSE values when we go from including zero variables, to only the first variable, to only the first two variables,...



Confidence intervals for predictions

- When making a confidence interval, it is important to be clear about the quantity that you are constructing a confidence interval for
- Once we estimate $\boldsymbol{\beta}$ we can construct intervals for various quantities
  - The average outcome for subjects with predictors $\boldsymbol{x}_{new}$
  - A prediction for a particular subject with predictors $\boldsymbol{x}_{new}$
- The difference between these has to do with reducible versus irreducible error

Confidence intervals for predictions

- The average outcome for subjects with predictors $\boldsymbol{x}_{new}$ is given by

$$E(Y|\boldsymbol{X} = \boldsymbol{x}_{new}) = \boldsymbol{x}_{new}\boldsymbol{\beta}$$

- A prediction for a particular subject with predictors $\boldsymbol{x}_{new}$ is given by

$$Y_{new} = \boldsymbol{x}_{new}\boldsymbol{\beta} + \epsilon$$

- No matter how much data we have, we can not reduce $\text{Var}(\epsilon) = \sigma^2$
- Confidence intervals for predictions of specific subjects are therefore wider than those for averages

Categorical predictors

- Many times our predictors include variables that are categorical
- How we include them into the model differs from quantitative variables
- Suppose we have a predictor X_j that denotes eye color
  - Assume only 3 levels: brown, green, and blue
- We can't simply include X_j into the model with $\beta_j X_j$

Categorical predictors

- We can instead include dummy variables
  - Need 1 less dummy variable than number of categories
- Define

$$I_1 = \begin{cases} 1, & X_j = \text{green} \\ 0, & \text{o/w} \end{cases} \qquad I_2 = \begin{cases} 1, & X_j = \text{blue} \\ 0, & \text{o/w} \end{cases}$$

- Then the regression model (if we only have X_j) becomes

$$E(Y|X_j) = \beta_0 + \beta_1 I_1 + \beta_2 I_2$$

Categorical predictors

- This implies that

$$E(Y|X_j) = \begin{cases} \beta_0, & X_j = \text{brown} \\ \beta_0 + \beta_1, & X_j = \text{green} \\ \beta_0 + \beta_2, & X_j = \text{blue} \end{cases}$$

- β_1 is interpreted as the average difference in the outcome between green and brown eyed subjects
  - Similar interpretation for β_2
- Brown is considered the baseline in the model
  - Choice of baseline does not affect model fit
  - Does change interpretation and magnitude of specific parameters

Categorical predictors

- If we want to test whether X_j is important, we must test

$$H_0 : \beta_1 = \beta_2 = 0$$

- Not common to test only one of these parameters at a time
- One nice feature of categorical covariates is that we don't have to make as many modeling decisions about how to include them in the model
    - Linear or nonlinear terms

Removing linear model assumptions

- There are two key assumptions that our models so far have generally made
  - Additivity
  - Linearity
- Additivity is the idea that the effect of X_j on the outcome does not depend on the levels of all other covariates
  - Not realistic in certain settings
- Linearity is simply when we assume the relationship between X_j and Y is linear
  - Also potentially problematic

Removing additivity

- In a linear model, the easiest way to remove additivity is through an interaction term
- Suppose we only have two covariates, X_1 and X_2
- The additive linear model assumes

$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Instead we can use the following model:

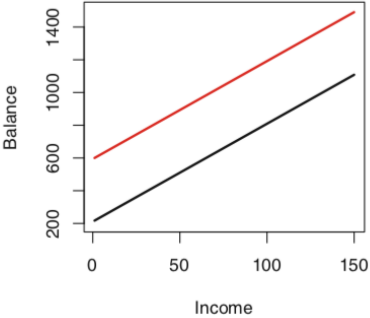
$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

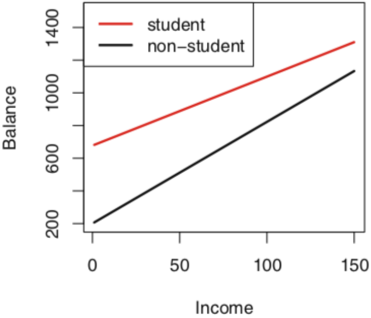
Removing additivity

- In the first model, a one unit change in X_1 is expected to lead to a change of β_1 in the outcome
  - Regardless of the value of X_2
- In the second model, a one unit change in X_1 is expected to lead to a change of $\beta_1 + \beta_3 X_2$ in the outcome
- The change now depends on X_2

Removing additivity

- If X_2 is binary, we can easily visualize this change
- The textbook has an example that tries to predict someone's bank balance given their income (quantitative) and student status (binary)
- The left plot is when we assume additivity, and the right plot is when we include an interaction





Removing linearity assumption

- What if instead we want to remove the linearity assumption?
- The easiest way is to include polynomial terms for X_j in the model
- Assume for now, we only have one covariate X
- The linear model assumes

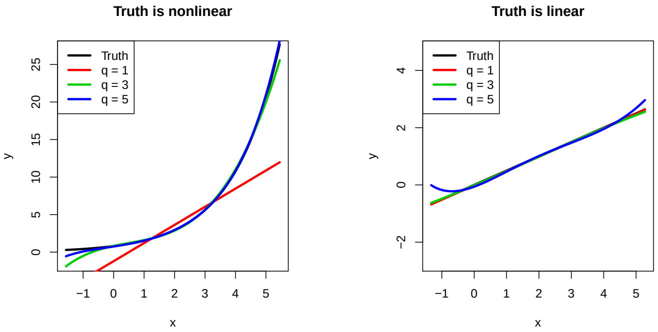
$$E(Y|X) = \beta_0 + \beta_1 X$$

- A polynomial model assumes

$$E(Y|X) = \beta_0 + \sum_{j=1}^{q} \beta_j X^j$$

Removing linearity assumption

- This allows for a much wider range of relationships between X and Y
- Let's investigate two scenarios and see how polynomial regression fares
  - Linear and nonlinear relationships
  - Vary degree q of the polynomial



Removing additivity and linearity assumptions

- We see that even in the linear model framework, we can somewhat alleviate problems caused by these two assumptions
- All of the approaches we considered above are still linear models
  - Just with the correct terms included
- Some of the models we will see later in class naturally account for these issues without having to manually specify them
  - Flexible, machine learning approaches

Other possible issues with linear models

- There are issues that could break our linear model assumptions
  - Correlated data
  - Non-constant variance
  - Outliers and high-leverage points
- We will not go into these in this class, but know they exist
- These are issues that are covered heavily in linear regression classes
  - Our textbook briefly mentions them and their possible fixes, but does not go into any detail