Financial Data Science

Lecture 3
Linear Regression in Finance

Liu Peng
liupeng@smu.edu.sg

Introducing linear regression https://youtu.be/n61tkVF6uAU
Deriving closed-form solution https://youtu.be/-H2hFOFjfoE

Probability Basics

- Probability distribution provides a framework for understanding and predicting the behavior of random variables
- Once we know the underlying data-generating probability distribution, we can make more informed decisions about how things are likely to appear, either in a predictive or optimization context.
- A probability distribution takes the random variable and converts it into a probability, which a floating number valued between 0 and 1.

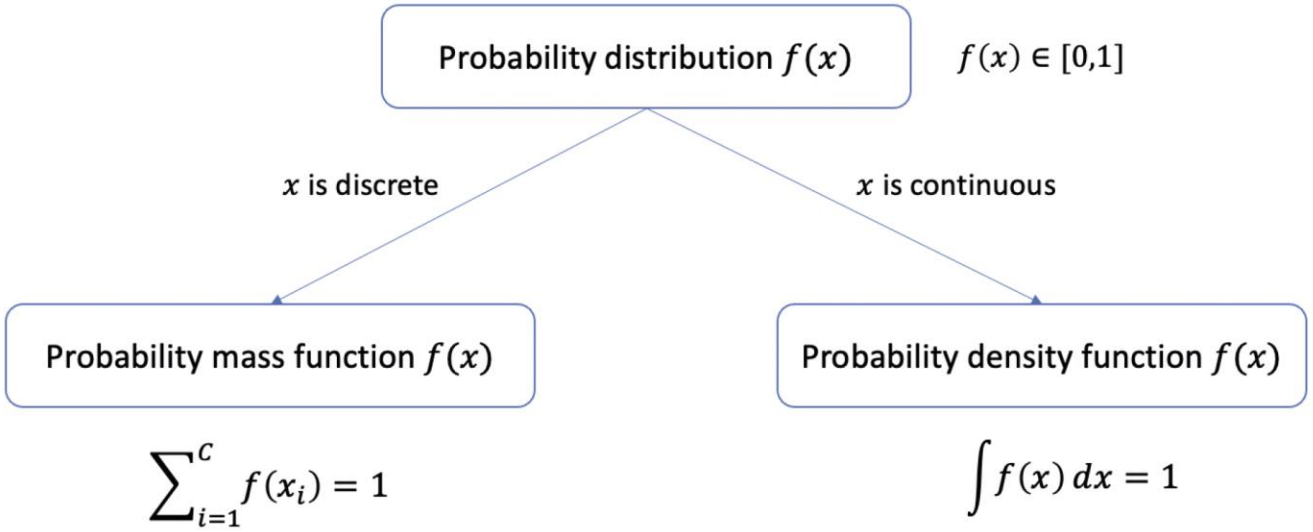


Figure 10.1 – Summarizing the two categories of probability distributions. Both distributions sum to 1

Random Variables and Probability Distributions

	Discrete probability distribution	Continuous probability distribution
Type of random variable	Distinct, countable, and separate, such as the number of heads in a series of coin tosses.	Can take any value within a continuous range, such as the weight of a person.
Probability representation	Use the PMF to describe the probability of each specific outcome or a range of outcomes, where probabilities are assigned to individual points.	Use the PDF to describe the probabilities of observing values within intervals or ranges. Probabilities are assigned to intervals, and the probability of any specific value is typically zero since the possible outcomes are uncountable.
Probability calculation	Probabilities are calculated for individual outcomes or a range of outcomes by summing the probabilities of the respective outcomes.	Probabilities are calculated for intervals or ranges of values by integrating the PDF over the desired range.
Visualization	Bar plots, where each bar represents the probability of a specific outcome or a range of outcomes.	Smooth curves where the area under the curve over a specific range represents the probability of the random variable falling within that range.

Figure 10.9 – Summarizing the differences between discrete and continuous probability distributions

Common Discrete Distributions

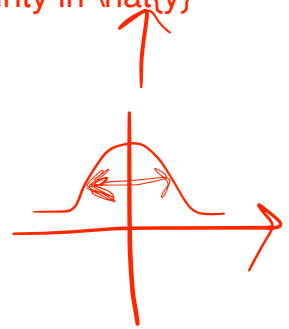
	Binomial distribution	Poisson distribution	Geometric distribution
Usage	Models the number of successes in a fixed number of independent Bernoulli trials with a constant probability of success	Models the number of events occurring in a fixed interval of time or space, given a constant average rate of occurrence	Models the number of trials required for the first success in a sequence of independent Bernoulli trials, each with the same probability of success
Parameters	n (number of trials) and p (probability of success)	λ (average rate of occurrence)	p (probability of success)
PMF	$P(x = k) = C(n, k)p^k(1-p)^{n-1}$	$P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$	$P(x = k) = (1-p)^{k-1}p$
Mean	$\mu = np$	$\mu = \lambda$	$\mu = \frac{1}{p}$
Variance	$\sigma^2 = np(1-p)$	$\sigma^2 = \lambda$	$\sigma^2 = \frac{1-p}{p^2}$
Applications	Coin tosses, quality control, opinion polling	Phone call arrivals, email arrivals, accidents at an intersection	Waiting time until an event occurs, number of attempts to achieve a desired outcome

Figure 10.8 – Summarizing and comparing different discrete distributions

What can we infer from a probability distribution?

uncertainty in beta0, beta1 —> uncertainty in \hat{y}

fixed x



In-class Quiz

- Q1-3

Linear Regression

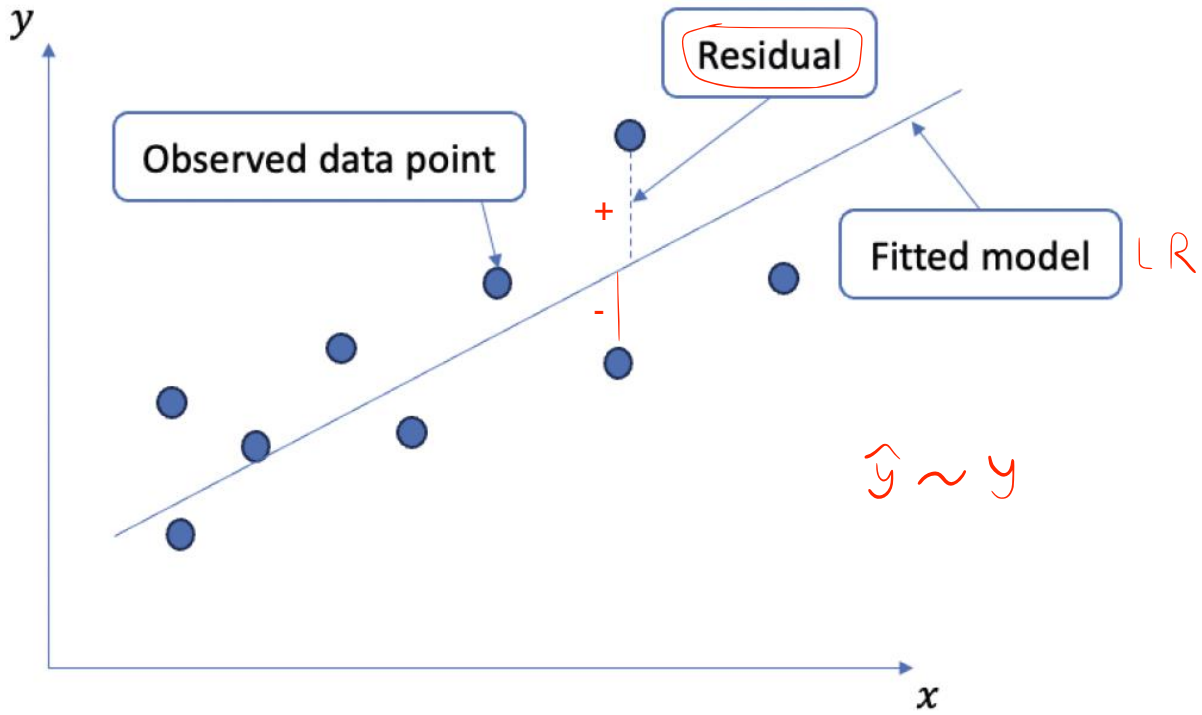
- At the core of linear regression is the concept of fitting a straight line, or more generally, a hyperplane, to the data points.
- Such fitting aims to minimize the deviation between the observed and predicted values.
- Obtaining an optimal model means identifying the best coefficients that define the relationship between the target variable and the input predictors.
- The expected outcome is modeled as a weighted sum of all the input variables

y

$\hat{y}$

frequentist

Visualizing LR Model



LR

$\hat{y} \sim y$

change {beta0, beta1}

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$\min SSR = \min \sum_{i=1}^{n} u_i^2 = \min (y_i - \hat{y}_i)^2$$

evaluate model

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

fixed

Var(y)

adjusted R^2

Q: what is the range of R2?

$(-\infty, 1]$

Figure 12.1 – The SLR model, where the linear model appears as a line and is trained by minimizing the SSR

Multiple Linear Regression

p features

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- MLR expands the single predictor in SLR to predict the target outcome based on multiple predictor variables
- A coefficient represents the change in the outcome variable for a single unit change in the associated predictor variable, assuming all other predictors are held constant.
- This is particularly useful in fields where the outcome variable is likely influenced by more than one predictor variable.

$$\hat{y}_{new} = \beta_0 + \beta_1 x_1 + \ldots + \beta_j \left(x_j + \Delta x_j \right) + \ldots + \beta_p x_p$$

$$\hat{y}_{old} = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_p x_p$$

$$\Delta\hat{y} = \hat{y}_{new} - \hat{y}_{old} = \beta_j \Delta x_j$$

$$\text{Adjusted } R^2 = 1 - \left(1 - R^2\right) \frac{(n-1)}{n-p-1}$$

\hat{y} = beta0 + beta1 x1 + beta2 x2

confounding variable

Simpson's Paradox

x1 = f(x2)

Simpson's Paradox says that a trend appears in different data groups but disappears or changes when combined.

Simpson's Paradox can appear when a variable that seems positively correlated with the outcome might be negatively correlated when we control other variables.

This paradox illustrates the importance of considering confounding variables and not drawing conclusions from aggregated data without understanding the context.

In-class Quiz

- Q4-6

Working with Categorical Variables

- In MLR, the process of including a binary predictor is similar to including a numeric predictor.
- However, the interpretation differs.

main effect

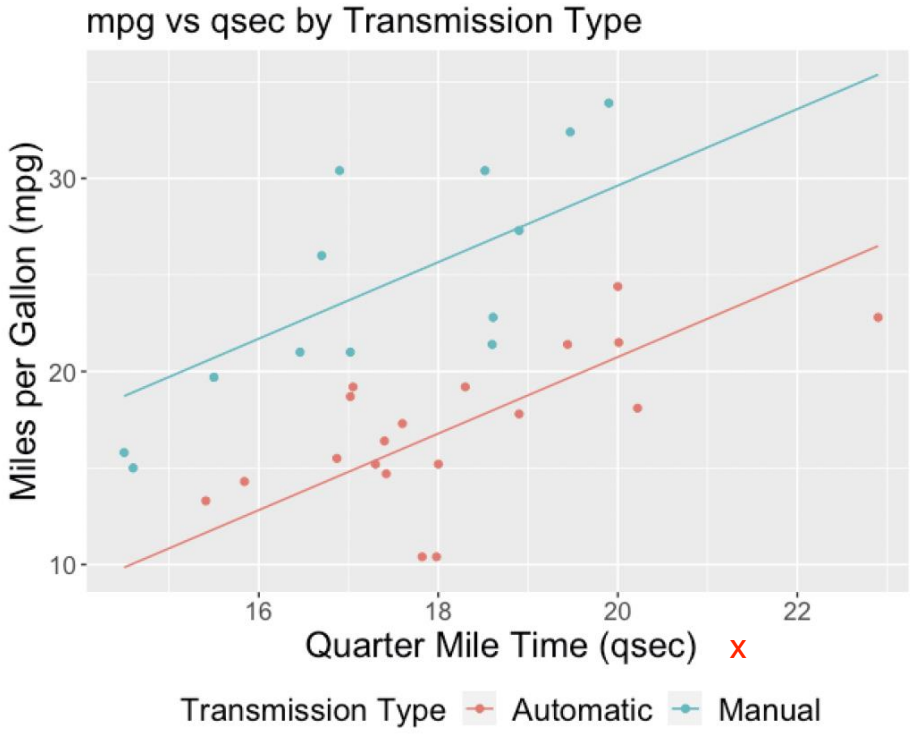
{0, 1}

interaction

$$\hat{y} = \beta_0 + \beta_1 x_{qsec} + \beta_2 x_{am\_cat}$$

$$\hat{y} = \beta_0 + \beta_1 x_{qsec} \quad 0$$

$$\hat{y} = \beta_0 + \beta_1 x_{qsec} + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_{qsec} \quad 1$$



y

x

Figure 12.3 – Visualizing the two linear regression models based on different transmission types. These two lines are parallel to each other due to a shift in the intercept term

Introducing the Interaction Term

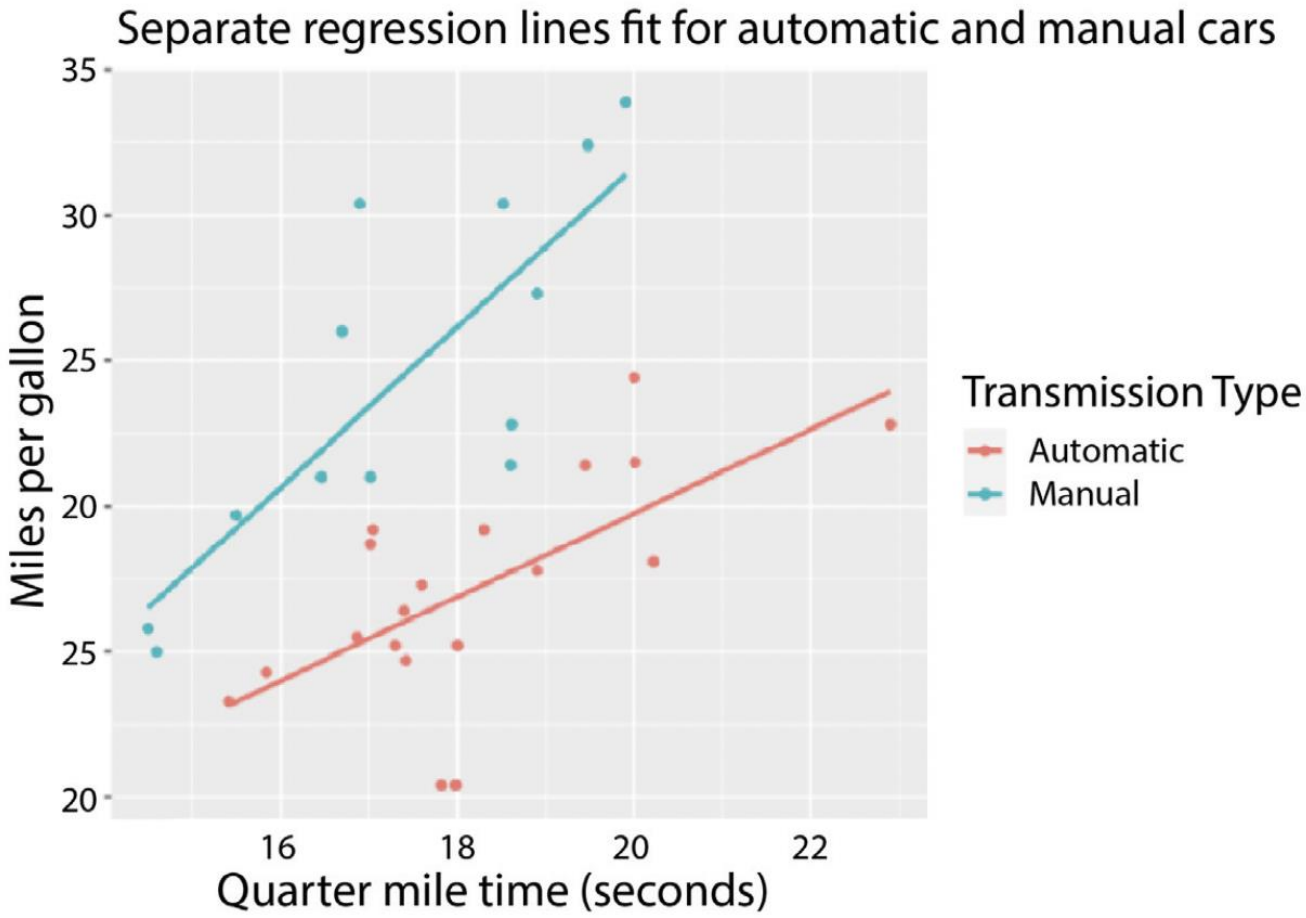


Figure 12.4 – Two intersecting lines due to the intersection term between quarter-mile time and transmission type

Introducing Nonlinear Terms

nonlinear terms introduce nonlinearity

linear model in coefficients

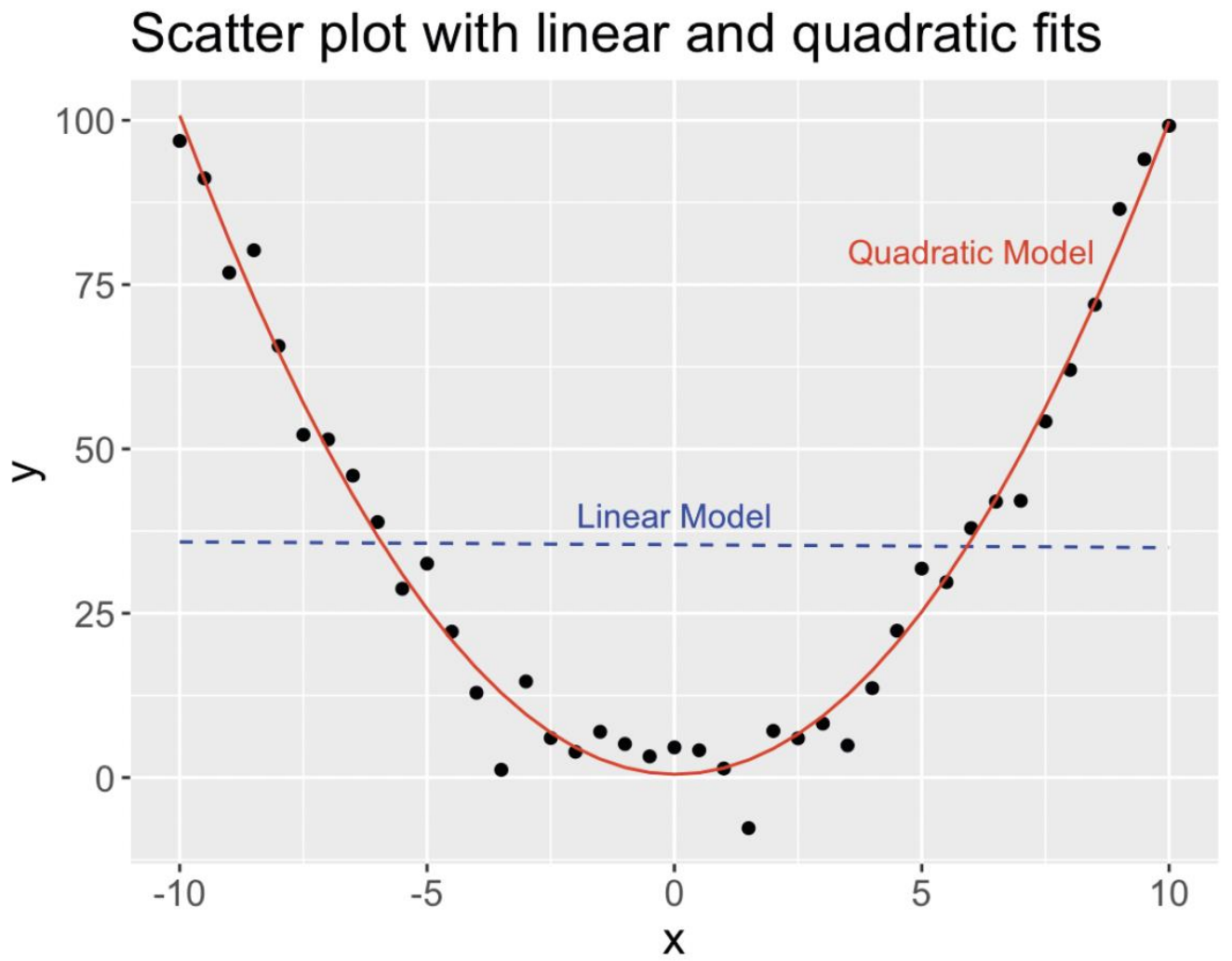


Figure 12.5 – Visualizing the linear and quadratic fits to the nonlinear data

Introducing Logarithmic Transformation

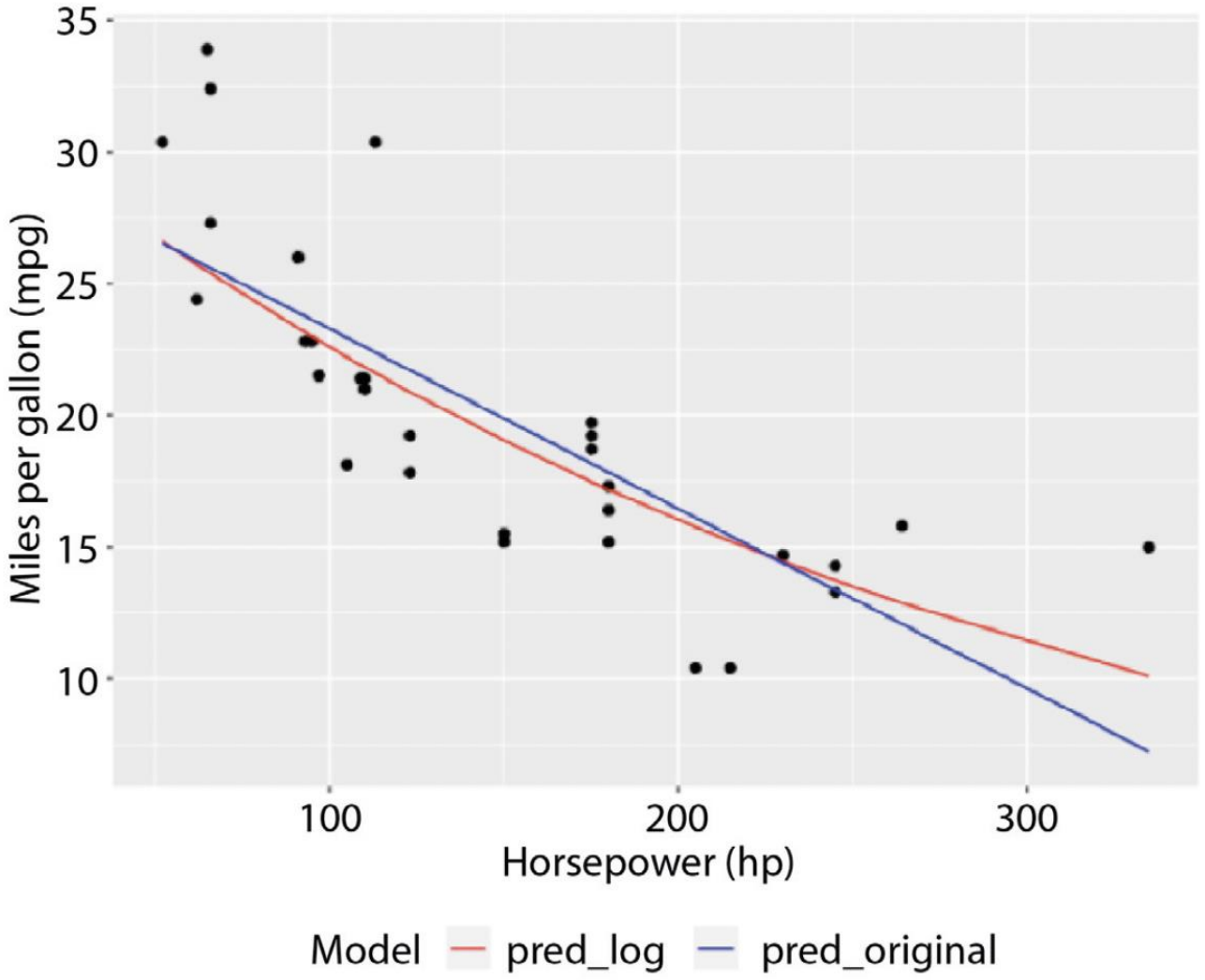


Figure 12.6 – Visualizing the original and log-transformed model

In-class Quiz

- Q7-9

Working with Closed-Form Solution

minimize $(y - X\beta)^T(y - X\beta)$

$$\frac{\partial(y^T y - 2\beta^T X^T y + \beta^T X^T X\beta)}{\partial\beta} = -2X^T y + 2X^T X\beta = 0$$

$$X^T X\beta = X^T y$$

$$\beta = (X^T X)^{-1} X^T y$$

- When developing a linear regression model, the available training set is given, and the only unknown parameters are the coefficients.
- It turns out that the closed-form solution to a linear regression model can be derived using the concept of the ordinary least squares (OLS) estimator, which aims to minimize the sum of the squared residuals in the model.
- Having the closed-form solution means we can simply plug in the required elements (in this case, X and y) and perform the calculation to obtain the solution, without resorting to any optimization procedure.

Ridge Regression

- Purely minimizing the RSS would give us an overfitting model, as represented by the high magnitude of the resulting coefficients.
- As a remedy, we could apply ridge regression by adding a penalty term to this loss function.
- Ridge regression, also referred to as L2 regularization, is a commonly used technique to alleviate overfitting in linear regression models by penalizing the magnitude of the estimated coefficients in the resulting model.

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

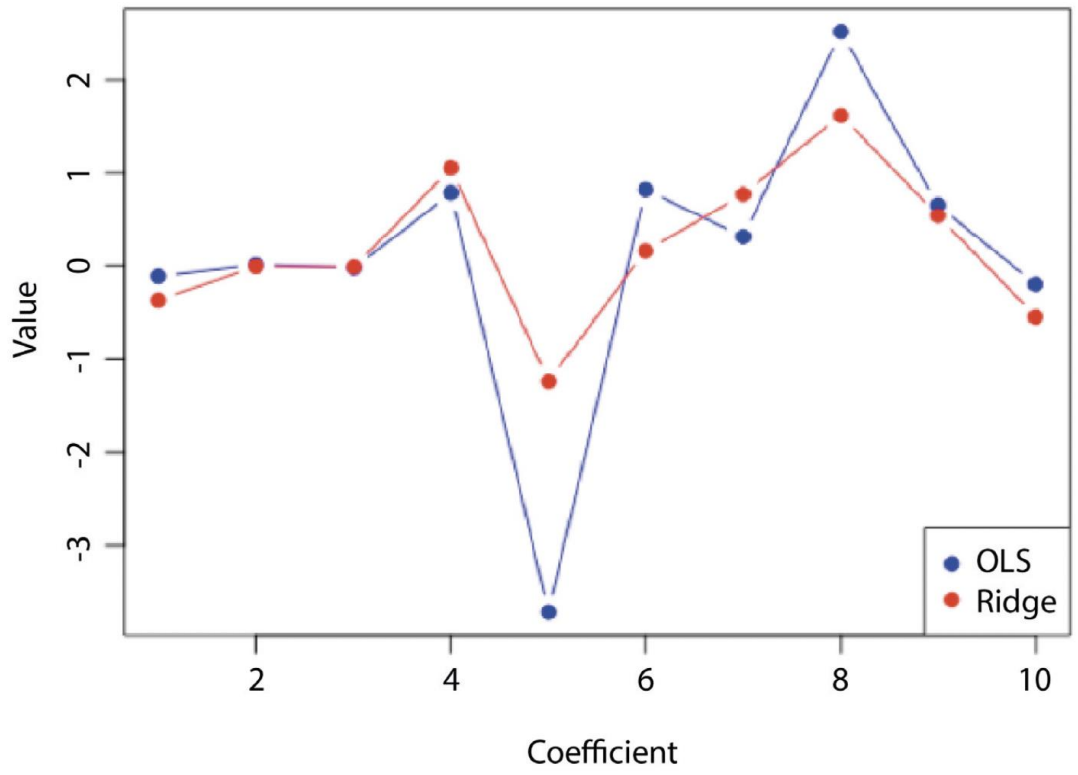


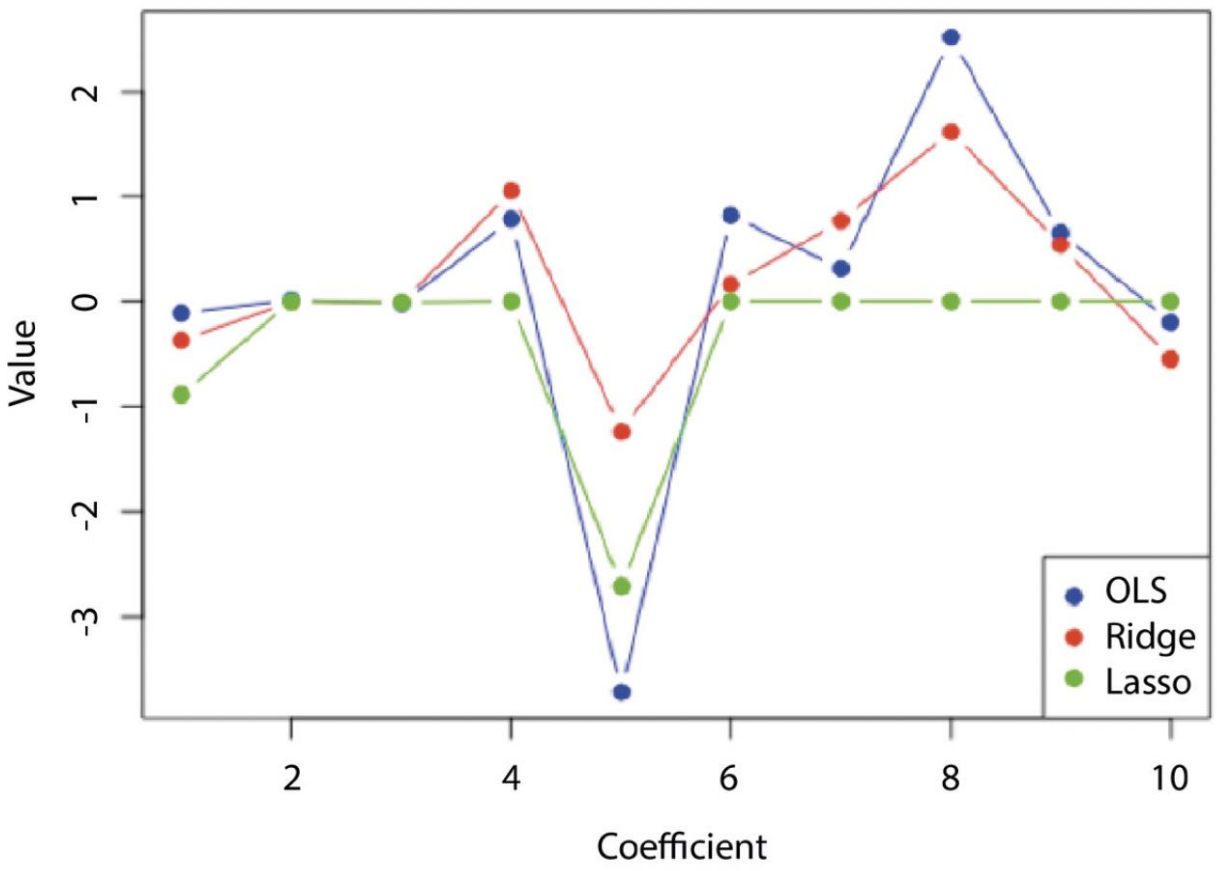
Figure 12.7 – Visualizing the estimated coefficients from the ridge and OLS models

Lasso Regression

- Lasso is similar to ridge regression but differs in terms of the specific process of calculating the magnitude of the coefficients.
- It uses the L1 norm of the coefficients, which consists of the total sum of absolute values of the coefficients, as the penalty that's added to the OLS loss function.
- The key characteristic of lasso regression is that it can reduce some coefficients exactly to 0, effectively performing variable selection.

sparse

$$L_{lasso} = RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$



In-class Quiz

- Q10-13

Reading materials

- Chapter 12, The Statistics and Machine Learning with R Workshop

Group Homework

- Predicting stock index returns
  - Choose S&P 500 index (SPY)
  - Predict its next-month monthly return for a period of 5 years (out of sample), using a rolling estimation window of past 5 years (in sample)
  - Evaluate prediction performance
  - Use different factor models (Fama-French 3 factors, 5 factors, etc.)
  - Use different linear models (LR, ridge, lasso)

Homework

- Second group homework to submit by one day before class starts next week
- Post learning reflections and questions in the group chat if any
- Review course contents and recording