Lecture 4: Linear Regression

LSE ME314: Introduction to Data Science and Machine Learning (https://github.com/me314-lse)

2025-07-17

Daniel de Kadt

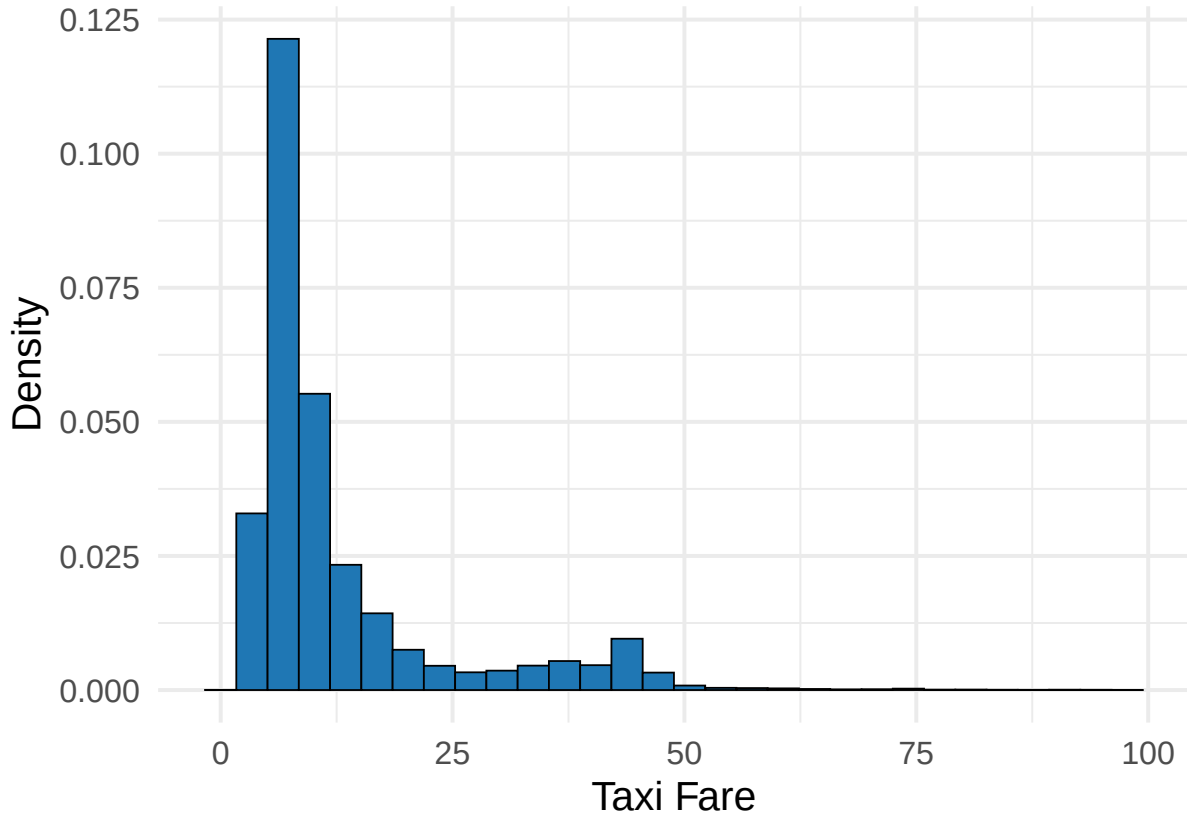
Slides last updated: 17 July 2025

Windy City Taxis

Source: John Lee/Chicago Tribune, Chicago Taxi Industry

Windy City Taxis

We're going to work with (a 5% sample of) 1m taxi trips in Chicago, from January 2016. Our outcome variable (aka dependent variable or response variable) will be the fare paid in USD:



Estimands vs. Estimators vs. Estimates

The DGP and Estimands

Let's back away from the observed data for a second.

Suppose we assume the following DGP for Y:

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

where β_0 and β_1 are fixed parameters, and $\epsilon \sim \mathcal{N}(0, 1)$.

The DGP and Estimands

Let's back away from the observed data for a second.

Suppose we assume the following DGP for Y:

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

where β_0 and β_1 are fixed parameters, and $\epsilon \sim \mathcal{N}(0, 1)$.

Remember: This is an *assumed* DGP for Y. It's probably wrong in reality.

The DGP and Estimands

Let's back away from the observed data for a second.

Suppose we assume the following DGP for Y:

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

where β_0 and β_1 are fixed parameters, and $\epsilon \sim \mathcal{N}(0, 1)$.

Remember: This is an *assumed* DGP for Y. It's probably wrong in reality.

Now connect this back to the CEF of Y. Given the assumed DGP, our *estimands* are:

- β_0: the intercept, $\mathbb{E}[Y|X = 0]$.
- β_1: the slope, $\frac{\partial \mathbb{E}[Y|X]}{\partial X}$.

Estimators

We have a sample with which we can attempt to learn about these parameters and the CEF.

Estimators

We have a sample with which we can attempt to learn about these parameters and the CEF.

The function we select to connect observed Y to X is our **estimator**. There are many ($\sim$ infinite) options:

Estimators

We have a sample with which we can attempt to learn about these parameters and the CEF.

The function we select to connect observed Y to X is our **estimator**. There are many ($\sim$ infinite) options:

- $\widehat{\mathbb{E}[Y|X]} = 42$

Estimators

We have a sample with which we can attempt to learn about these parameters and the CEF.

The function we select to connect observed Y to X is our **estimator**. There are many ($\sim$ infinite) options:

- $\widehat{\mathbb{E}[Y|X]} = 42$
- $\widehat{\mathbb{E}[Y|X]} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Estimators

We have a sample with which we can attempt to learn about these parameters and the CEF.

The function we select to connect observed Y to X is our **estimator**. There are many ($\sim$ infinite) options:

- $\widehat{\mathbb{E}[Y|X]} = 42$
- $\widehat{\mathbb{E}[Y|X]} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- $\widehat{\mathbb{E}[Y|X]} = \text{median}(Y)$

Estimators

We have a sample with which we can attempt to learn about these parameters and the CEF.

The function we select to connect observed Y to X is our **estimator**. There are many ($\sim$ infinite) options:

- $\widehat{\mathbb{E}[Y|X]} = 42$
- $\widehat{\mathbb{E}[Y|X]} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- $\widehat{\mathbb{E}[Y|X]} = \text{median}(Y)$
- $\widehat{\mathbb{E}[Y|X]} = \hat{\beta}_0 + \hat{\beta}_1 X \longrightarrow$ 'linear regression estimator'

Estimators

We have a sample with which we can attempt to learn about these parameters and the CEF.

The function we select to connect observed Y to X is our **estimator**. There are many ($\sim$ infinite) options:

- $\widehat{\mathbb{E}[Y|X]} = 42$
- $\widehat{\mathbb{E}[Y|X]} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- $\widehat{\mathbb{E}[Y|X]} = \text{median}(Y)$
- $\widehat{\mathbb{E}[Y|X]} = \hat{\beta}_0 + \hat{\beta}_1 X \longrightarrow$ 'linear regression estimator'

Once we combine any of these with our particular sample, we get an **estimate**.

Properties of Estimators

Recall the central limit theorem (CLT): Over repeated samples from a random variable, the distribution of sample means is approximately normal.

This distribution is called the **sampling distribution** of the random variable.

Properties of Estimators

Recall the central limit theorem (CLT): Over repeated samples from a random variable, the distribution of sample means is approximately normal.

This distribution is called the **sampling distribution** of the random variable.

It turns out that estimators are themselves random variables – they are functions of the observed data drawn from the probabilistic DGP.

Properties of Estimators

We can evaluate estimators in terms of two properties of their sampling distribution:

- **Bias**: The difference between the expected value of the estimator and the true value of the parameter. An estimator is unbiased if its expected value equals the true parameter value.
  - E.g. does $\mathbb{E}[\hat{\mu}] = \mu$? for the mean
  - E.g. does $\mathbb{E}[\hat{\beta}_1] = \beta_1$? for linear regression

Properties of Estimators

We can evaluate estimators in terms of two properties of their sampling distribution:

- **Bias**: The difference between the expected value of the estimator and the true value of the parameter. An estimator is unbiased if its expected value equals the true parameter value.
  - E.g. does $\mathbb{E}[\hat{\mu}] = \mu$? for the mean
  - E.g. does $\mathbb{E}[\hat{\beta}_1] = \beta_1$? for linear regression
- **Efficiency**: The variability of the estimator across repeated samples:
  - E.g. compare the variance of two proposed estimators: $\text{Var}(\hat{\mu}_1)$ vs. $\text{Var}(\hat{\mu}_2)$

Variance

Yesterday we covered the **expected value** of a random variable – average value of that variable over repeated samples.

A related concept is the **variance** of a random variable – the spread of that variable over repeated samples.

Variance

Yesterday we covered the **expected value** of a random variable – average value of that variable over repeated samples.

A related concept is the **variance** of a random variable – the spread of that variable over repeated samples.

The variance of a random variable (which could be a function!) X is defined as:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Read this as: The difference between the random variable and its expected value, squared.

Variance

Yesterday we covered the **expected value** of a random variable – average value of that variable over repeated samples.

A related concept is the **variance** of a random variable – the spread of that variable over repeated samples.

The variance of a random variable (which could be a function!) X is defined as:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Read this as: The difference between the random variable and its expected value, squared.

Intuition check: What is the variance of $\widehat{\mathbb{E}[Y|X]} = 42$?

Variance

Yesterday we covered the **expected value** of a random variable – average value of that variable over repeated samples.

A related concept is the **variance** of a random variable – the spread of that variable over repeated samples.

The variance of a random variable (which could be a function!) X is defined as:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Read this as: The difference between the random variable and its expected value, squared.

Intuition check: What is the variance of $\widehat{\mathbb{E}[Y|X]} = 42$?

Zero, because the estimator is a constant so $X = \mathbb{E}[X]$ always.

Bivariate Linear Regression

Assumed Model and Proposed Estimator

So far we have assumed the following DGP for Y:

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X + \epsilon$$

where β_0 and β_1 are fixed parameters.

Assumed Model and Proposed Estimator

So far we have assumed the following DGP for Y:

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X + \epsilon$$

where β_0 and β_1 are fixed parameters.

Let's estimate these parameters using the ordinary least squares (OLS) linear regression:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Remember, the function we estimate will be a straight line with some slope $\hat{\beta}_1$ and zero-intercept $\hat{\beta}_0$.

Ordinary Least Squares

The ordinary least squares line is the 'line of best fit'.

This is defined as the line that minimizes:

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i$ is the 'predicted value' of Y_i, given the observed value X_i.

SSR stands for **sum of squared residuals**, where $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ is the residual for i.

Ordinary Least Squares

The ordinary least squares line is the 'line of best fit'.

This is defined as the line that minimizes:

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i$ is the 'predicted value' of Y_i, given the observed value X_i.

SSR stands for **sum of squared residuals**, where $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ is the residual for i.

We call this the **objective function** of OLS.

Fitting the linear regression line is thus a minimization problem – there are many possible values of $\hat{\beta}_0$ and $\hat{\beta}_1$, but we need to pick the ones that satisfy this objective function.

Ordinary Least Squares: The Weeds

To see how to solve this minimization problem it's easiest if we work in matrix algebra. This **won't be on the test**!

Assume a $(n \times 1)$ **response vector**, and a $(n \times 2)$ **design matrix** that includes a single feature with an intercept (a column of 1s):

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{\text{Response vector } \mathbf{Y}} \qquad \underbrace{\begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}}_{\text{Design matrix } \mathbf{X}}$$

Ordinary Least Squares: The Weeds

Our estimators can likewise be represented as a vector:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

This is a 2×1 vector, given we want to estimate β_0 and β_1.

Ordinary Least Squares: The Weeds

Our estimators can likewise be represented as a vector:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

This is a 2×1 vector, given we want to estimate β_0 and β_1.

In matrix notation, we can rewrite our objective function as:

$$\hat{\beta} = \arg\min_{\beta} \| Y - X\beta \|^2$$

where $\| Y - X\beta \|^2 = (Y - X\beta)'(Y - X\beta)$.

Ordinary Least Squares: The Weeds

Step 1: Expand the objective function S

$$S(\hat{\beta}) = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

$$= \mathbf{Y}'\mathbf{Y} - 2\hat{\beta}'\mathbf{X}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}$$

Ordinary Least Squares: The Weeds

Step 1: Expand the objective function S

$$S(\hat{\beta}) = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

$$= \mathbf{Y}'\mathbf{Y} - 2\hat{\beta}'\mathbf{X}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}$$

Step 2: Take the first derivative with respect to $\hat{\beta}$

$$\frac{\partial S}{\partial \hat{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\beta}$$

Ordinary Least Squares: The Weeds

Step 1: Expand the objective function S

$$S(\hat{\beta}) = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

$$= \mathbf{Y}'\mathbf{Y} - 2\hat{\beta}'\mathbf{X}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}$$

Step 2: Take the first derivative with respect to $\hat{\beta}$

$$\frac{\partial S}{\partial \hat{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\beta}$$

Step 3: Set the gradient to zero

$$-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

Ordinary Least Squares: The Weeds

Step 1: Expand the objective function S

$$S(\hat{\beta}) = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

$$= \mathbf{Y}'\mathbf{Y} - 2\hat{\beta}'\mathbf{X}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}$$

Step 2: Take the first derivative with respect to $\hat{\beta}$

$$\frac{\partial S}{\partial \hat{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\beta}$$

Step 3: Set the gradient to zero

$$-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

Step 4: Solve for $\hat{\beta}$

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Y}$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

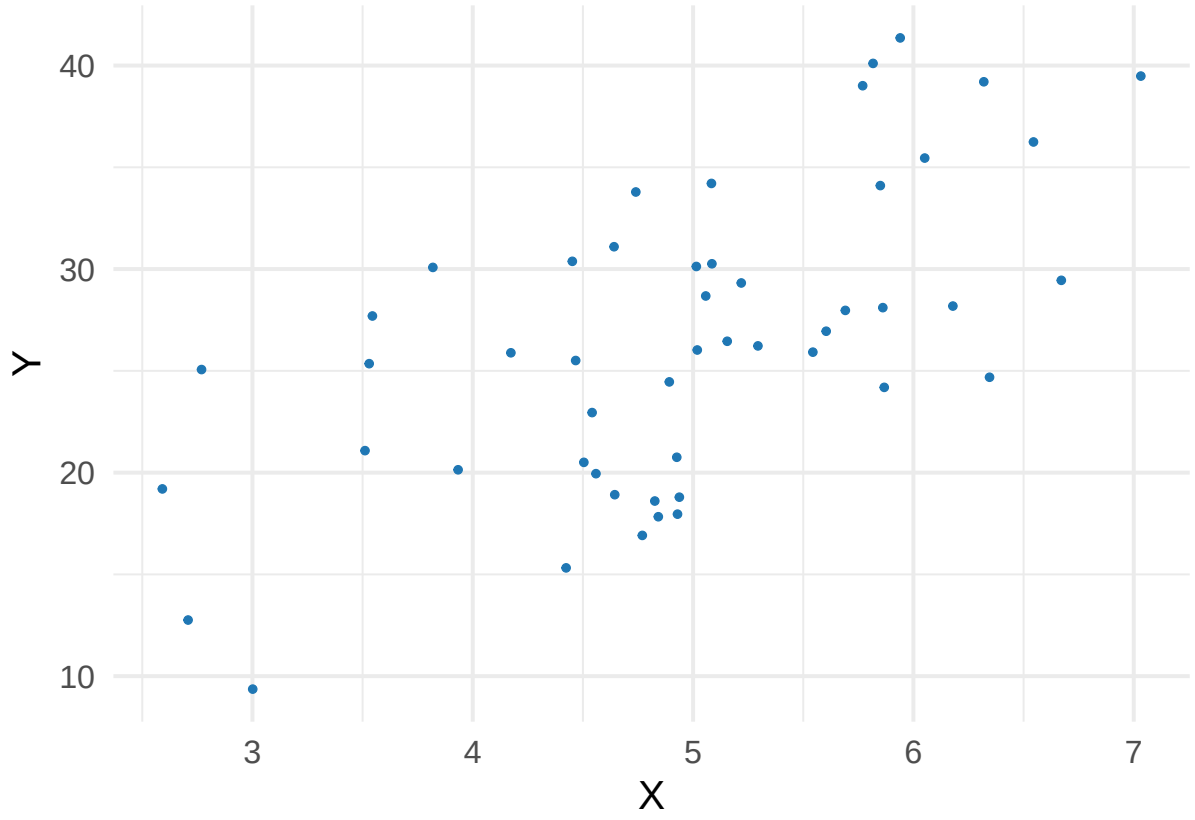
Ordinary Least Squares in Action

Whew!

Ordinary Least Squares in Action

Whew! Let's see how this works in practice:

```
beta_0 <- 2
beta_1 <- 5
X <- rnorm(50, mean = 5, sd = 1)
Y <- beta_0 + beta_1 * X + runif(50, -10, 10)
```



Ordinary Least Squares in Action

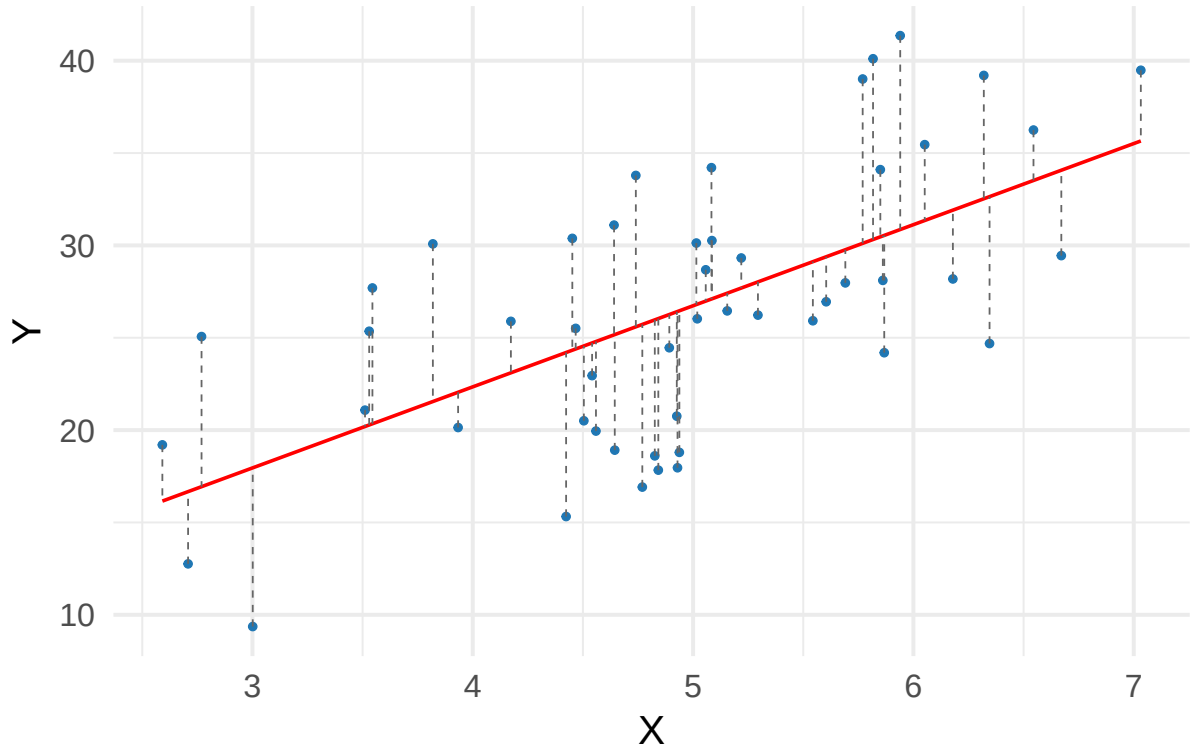
Now, we will solve for the OLS estimates of β_0 and β_1:

```
X_mat <- cbind(1, X)
beta_hat <- solve(t(X_mat) %*% X_mat) %*% t(X_mat) %*% Y
Y_hat <- X_mat %*% beta_hat
residuals <- Y - Y_hat
```

Ordinary Least Squares in Action

Now, we will solve for the OLS estimates of β_0 and β_1:

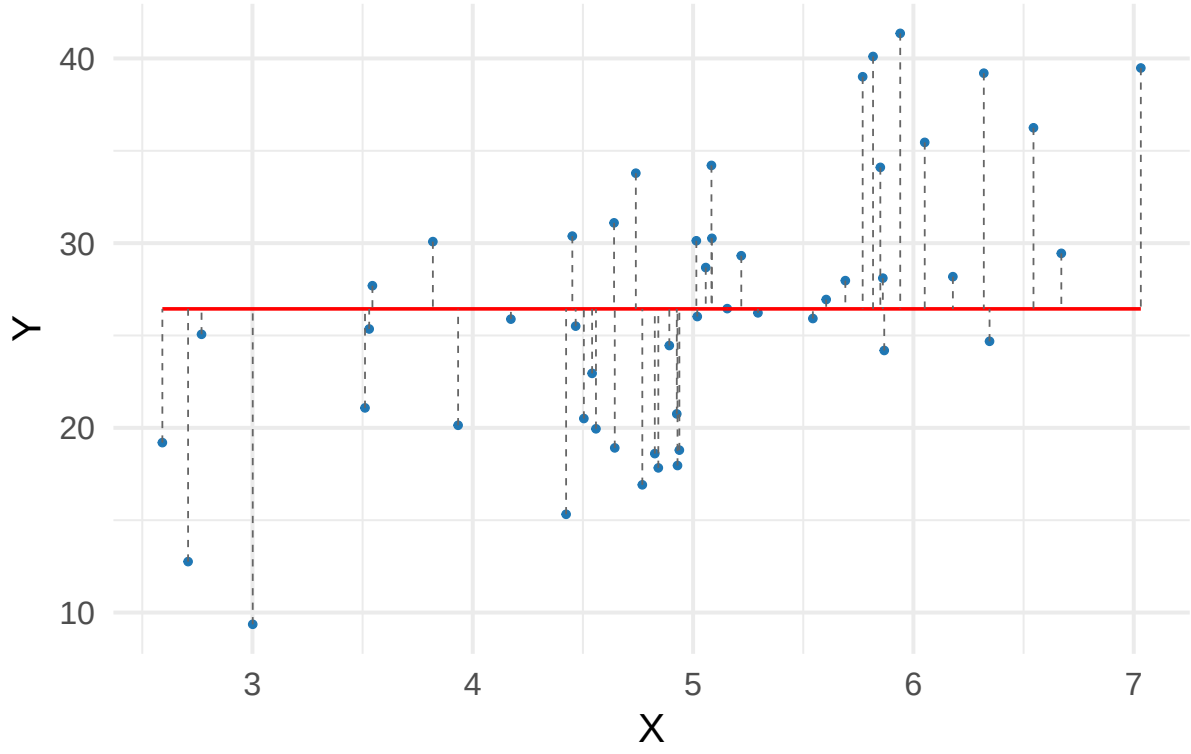
```
X_mat <- cbind(1, X)
beta_hat <- solve(t(X_mat) %*% X_mat) %*% t(X_mat) %*% Y
Y_hat <- X_mat %*% beta_hat
residuals <- Y - Y_hat
```



Ordinary Least Squares in Action

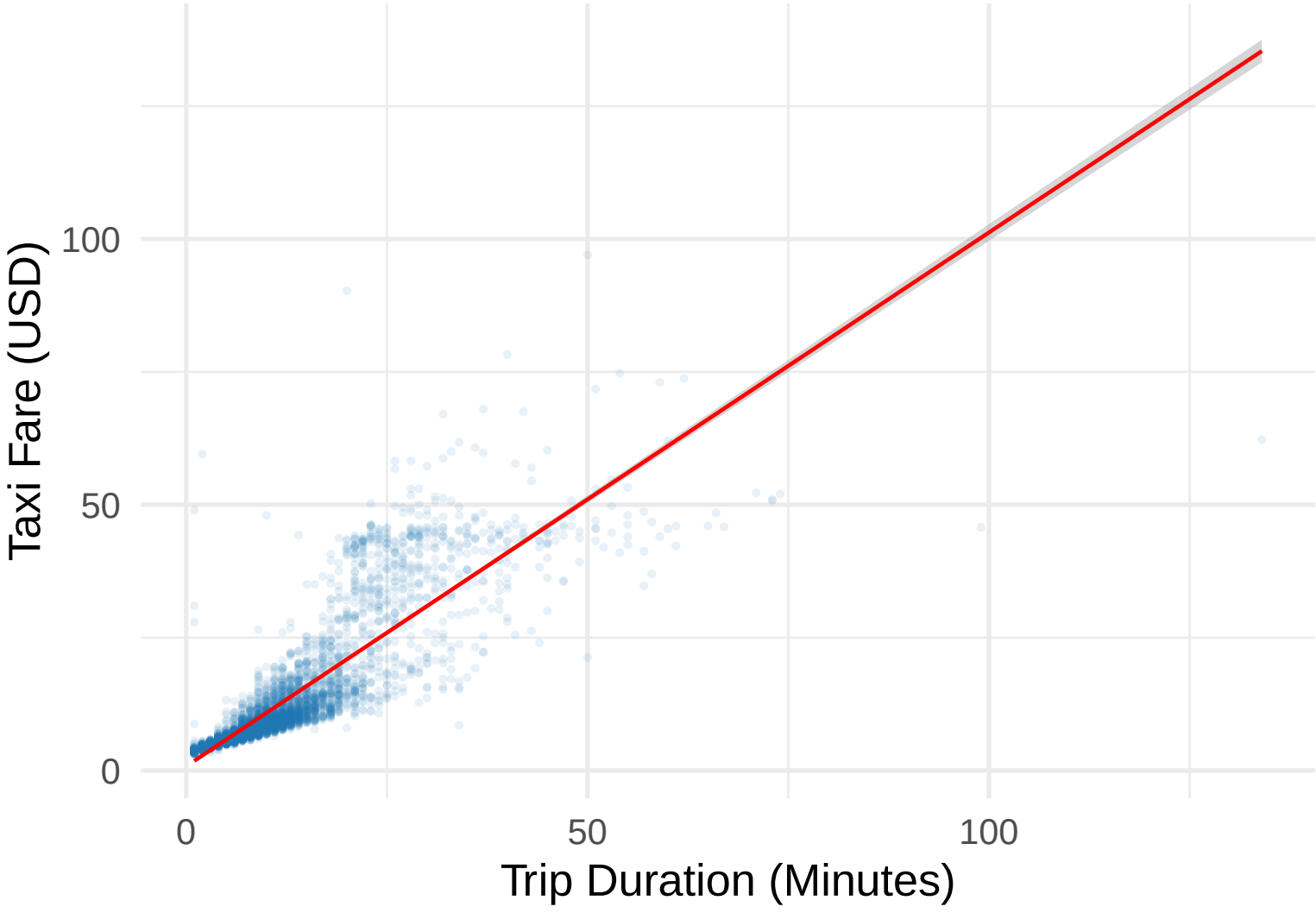
Contrast this performance with the 'null model' – where we explicitly assert that $\beta_1 = 0$:

```
Y_hat_null <- rep(mean(Y), 50)
residuals_null <- Y - Y_hat_null
```



Ordinary Least Squares in Action: Chicago

Let's take this to real data



Ordinary Least Squares in Action

```
lm(fare ~ trip_miles, data = taxi_trips) %>%
  summary() %>%
  broom::tidy()
```

```
# A tibble: 2 x 5
  term         estimate std.error statistic p.value
  <chr>           <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    12.5     0.0532       234.        0
2 trip_miles      0.103   0.00233       44.2       0
```

Interpretation:

- Intercept (β_0: The predicted value of Y when $X = 0$
- Slope (β_1): The predicted change in Y for a one unit change in X.

Interpretation: Marginal Changes

The slope gives us the marginal change or marginal effect.

This tells us how Y changes for a one unit change in X.

What this means in a given regression will depend on:

- The units of Y and X.
- The functional form of Y and X (log or not)

OLS Assumptions: Unbiasedness

Under what conditions is our OLS estimator an unbiased estimator of the parameters in the DGP (and thus the population CEF of Y)?

OLS Assumptions: Unbiasedness

Under what conditions is our OLS estimator an unbiased estimator of the parameters in the DGP (and thus the population CEF of Y)?

1. **Linearity**: The true relationship between Y and X is linear *in the parameters*
   - ➜ A violation: the relationship is not linear, e.g., $Y = \beta_0 + \beta_1{}^2 X + \epsilon$.

OLS Assumptions: Unbiasedness

Under what conditions is our OLS estimator an unbiased estimator of the parameters in the DGP (and thus the population CEF of Y)?

1. **Linearity**: The true relationship between Y and X is linear *in the parameters*
   - ➜ A violation: the relationship is not linear, e.g., $Y = \beta_0 + \beta_1{}^2 X + \epsilon$.

2. **I.i.d. random sample**: The sample is a random draw.
   - ➜ A violation: our sample over-represents certain values from the DGP.

OLS Assumptions: Unbiasedness

3. **No multicollinearity**: Features are not perfectly correlated with each other (this causes non-invertibility of the design matrix $\mathbf{X}$).
   - ➜ A violation: two features that are a linear combination of each other.

OLS Assumptions: Unbiasedness

3. **No multicollinearity**: Features are not perfectly correlated with each other (this causes non-invertibility of the design matrix $\mathbf{X}$).
   - ➜ A violation: two features that are a linear combination of each other.

4. **Zero conditional mean**: $E[\epsilon|X] = 0$ for all X.
   - ➜ A violation: the error term is correlated with X (e.g. confounding)
   - ➜ Note: This is a big one. It basically means, 'all the systematic components in the DGP are in our estimator'.

OLS Assumptions: Unbiasedness

3. **No multicollinearity**: Features are not perfectly correlated with each other (this causes non-invertibility of the design matrix $\mathbf{X}$).
   - ➜ A violation: two features that are a linear combination of each other.

4. **Zero conditional mean**: $E[\epsilon|X] = 0$ for all X.
   - ➜ A violation: the error term is correlated with X (e.g. confounding)
   - ➜ Note: This is a big one. It basically means, 'all the systematic components in the DGP are in our estimator'.

With all four assumptions met, OLS is an unbiased estimator of the parameters in the DGP (and thus the population CEF of Y).

Gauss-Markov Theorem

Consider a fifth assumption:

5. **Homoscedasticity**: $Var(\varepsilon|X) = \sigma^2$ for all X.
   - ➜ The variance of the error term is uncorrelated with the feature X
   - ➜ A violation: Y values in certain parts of the span of X are systematically further from the regression line than others.

Gauss-Markov Theorem

Consider a fifth assumption:

5. **Homoscedasticity**: $Var(\varepsilon|X) = \sigma^2$ for all X.
   - ➜ The variance of the error term is uncorrelated with the feature X
   - ➜ A violation: Y values in certain parts of the span of X are systematically further from the regression line than others.

If these five assumptions are met, then we say that linear regression is the **Best Linear Unbiased Estimator** (BLUE) of the (population) CEF of Y. (Best means most efficient here).

This is called the **Gauss-Markov Theorem**.

Gauss-Markov Theorem

Consider a fifth assumption:

5. **Homoscedasticity**: $Var(\varepsilon|X) = \sigma^2$ for all X.
   - ➜ The variance of the error term is uncorrelated with the feature X
   - ➜ A violation: Y values in certain parts of the span of X are systematically further from the regression line than others.

If these five assumptions are met, then we say that linear regression is the **Best Linear Unbiased Estimator** (BLUE) of the (population) CEF of Y. (Best means most efficient here).

This is called the **Gauss-Markov Theorem**.

Even when the assumptions are not met, linear regression can still have desirable properties, which is one of the reasons it is the workhorse of econometrics and data science.

Diagnostics: Classical

R^2: The proportion of variance in Y explained by the model.

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}}$$

where SSR is the sum of squared residuals (the unexplained variation in Y), and SST is the total sum of squares (all the variation in Y).

Intuition:

- If $R^2 = 0$, then the model explains none of the variation in Y.
- If $R^2 = 1$, then the model explains all of the variation in Y.

Does a high R^2 always make for a good model? (More later)

What Good is a Regression?*

So far we have assumed that we are using regression to estimate the parameters that define the DGP of Y.

In effect, we are using regression as a:

1. *Causal device*:
   - ➜ Estimate the causal effect of a feature on the outcome

What Good is a Regression?*

So far we have assumed that we are using regression to estimate the parameters that define the DGP of Y.

In effect, we are using regression as a:

1. *Causal device*:
   ➜ Estimate the causal effect of a feature on the outcome

But we could use regression in other ways:

2. As a *descriptive device*:
   ➜ Summarize a correlation or co-occurence between two variables, irrespective of any DGP
3. As a *predictive device*:
   ➜ Fill in 'missing values' (unseen realizations) of Y using X

*Spirling & Stewart (2023)

Multivariate Linear Regression

Setup

Let's expand our linear regression model by including more than one feature:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

where $\hat{\beta}_0$ is the intercept, $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are the slopes for each feature.

This is called **multivariate linear regression**.

Multivariate OLS: Matrix Representation

Note that with matrices we can neatly summarise the multivariate case by expanding the design matrix (and the parameter vector):

$$\underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\text{Estimators } \beta} \qquad \underbrace{\begin{bmatrix} 1 & X_1 & X_2 & \dots & X_k \\ 1 & X_1 & X_2 & \dots & X_k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_1 & X_2 & \dots & X_k \end{bmatrix}}_{\text{Design matrix } \mathbf{X}}$$

Now we can write down our multivariate OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Multivariate Linear Regression in Action

```
lm(fare ~ trip_mins + trip_miles, data = taxi_trips) %>%
  summary() %>%
  broom::tidy()
```

```
# A tibble: 3 x 5
  term         estimate std.error statistic   p.value
  <chr>           <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)    1.30     0.0427       30.5 6.29e-202
2 trip_mins      0.946    0.00273     347.  0
3 trip_miles     0.0291   0.00125      23.3 2.01e-119
```

Note that we now have two slopes, one for each feature.

Interpretation

In bivariate regression the slope for any one feature was easy to interpret: the change in Y for a one unit change in that feature.

In multivariate regression, we can think of the CEF of Y as an 'n-dimensional hyperplane' (helpful)

The slope for any one feature gives us the change in Y for a one unit change in that feature, holding all other features constant (*ceteris paribus*).

We call this the **partialling out** interpretation, and refer to the slopes as 'partial effects'

Interpretation: Partial Effects

What's going on here?

```
# Fit a bivariate regression of fare on miles, and save residuals
lm_miles <- lm(fare ~ trip_miles, data = taxi_trips)
taxi_trips$pred_fare_miles <- predict(lm_miles)
taxi_trips$res_fare_miles <- taxi_trips$fare - taxi_trips$pred_fare_miles

# Fit a bivariate regression of mins on miles, and save residuals
lm_mins <- lm(trip_mins ~ trip_miles, data = taxi_trips)
taxi_trips$pred_mins_miles <- predict(lm_mins)
taxi_trips$res_mins_miles <- taxi_trips$trip_mins - taxi_trips$pred_mins_miles

# Fit a regression of residuals on residuals
lm(res_fare_miles ~ res_mins_miles, data = taxi_trips) %>%
  summary() %>%
  broom::tidy()
```

```
# A tibble: 2 x 5
  term            estimate std.error statistic p.value
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    -1.65e-12   0.0278  -5.94e-11    1.00
2 res_mins_miles  9.46e- 1   0.00273  3.47e+ 2    0
```

More Complex Specifications

We can also include more complex specifications in our regression, such as:

- **Categorical variables**: Qualitative features that we will encode as dummy variables.
- **Polynomials**: Non-linear expansions of features, such as squares or cubes.
- **Interactions**: Flexibility that allows features to affect Y differently as a function of other features.

More Complex Specifications: Categorical Variables

```
# What are the levels of the categorical variable?
unique(taxi_trips$payment_type)
```

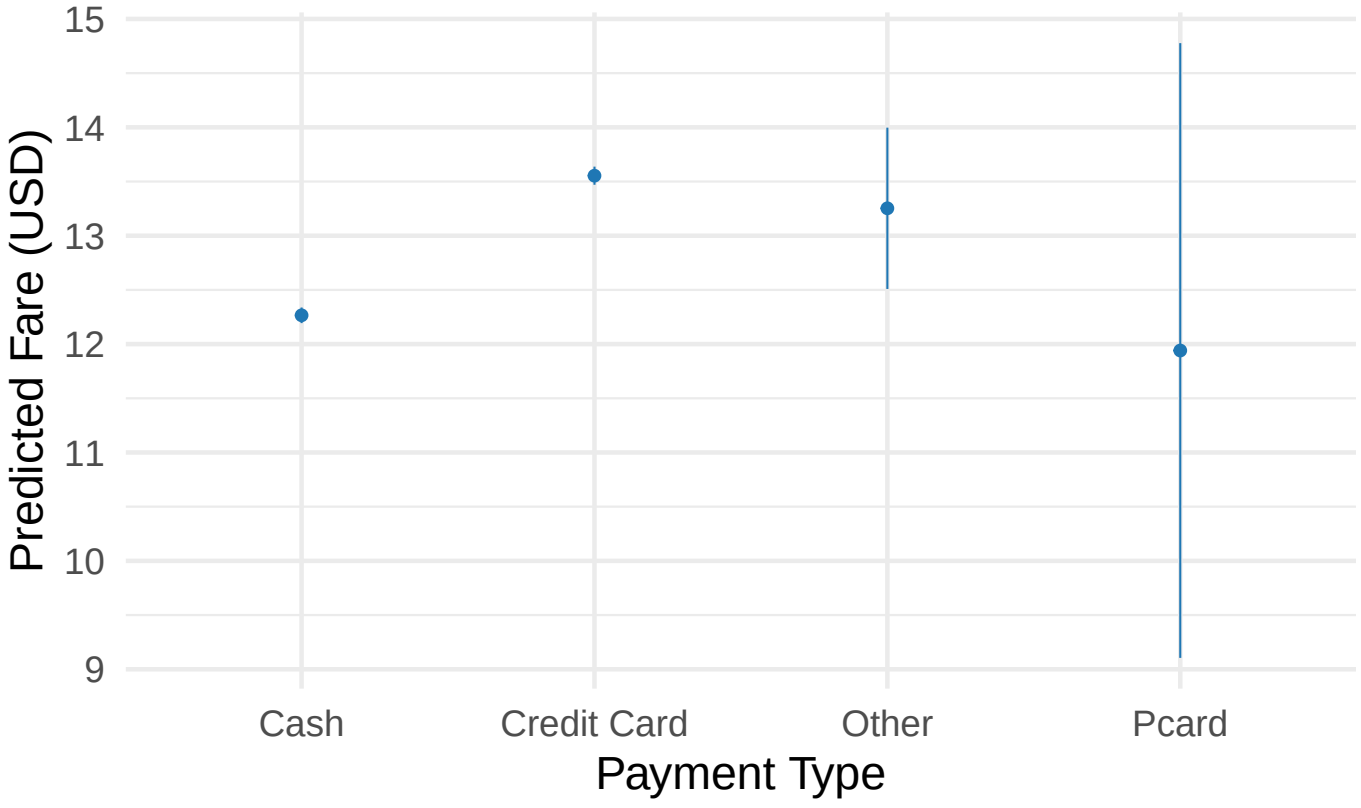
```
[1] "Cash"        "Credit Card" "Other"       "Pcard"
```

```
lm(fare ~ trip_mins + trip_miles + factor(payment_type),
   data = taxi_trips) %>%
  summary() %>%
  broom::tidy()
```

```
# A tibble: 6 x 5
  term                                estimate std.error statistic   p.value
  <chr>                                  <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)                            0.850    0.0469      18.1   4.55e- 73
2 trip_mins                              0.937    0.00273    343.    0
3 trip_miles                             0.0288   0.00124     23.2   1.31e-118
4 factor(payment_type)Credit Card        1.29     0.0566      22.8   3.86e-114
5 factor(payment_type)Other              0.987    0.381        2.59  9.63e-  3
6 factor(payment_type)Pcard             -0.325    1.45        -0.224 8.22e-  1
```

More Complex Specifications: Categorical Variables

Here we visualize the predicted fare paid based on different payment types, holding trip_mins and trip_miles constant at their means.



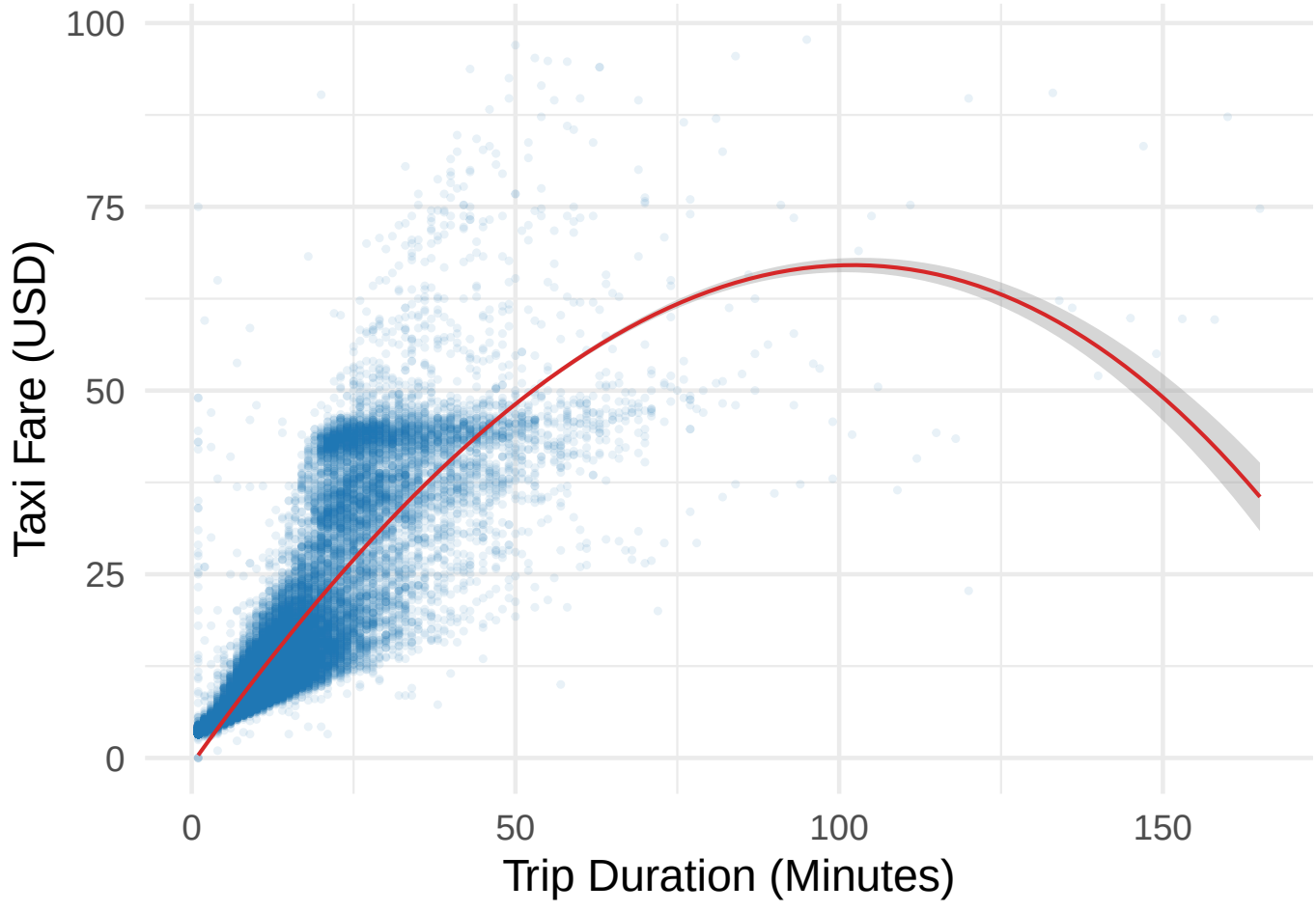
More Complex Specifications: Polynomials

```
lm(fare ~ trip_mins + I(trip_mins^2) + I(trip_mins^3),
   data = taxi_trips) %>%
  summary() %>%
  broom::tidy()
```

```
# A tibble: 4 x 5
  term              estimate  std.error statistic  p.value
  <chr>                <dbl>      <dbl>     <dbl>    <dbl>
1 (Intercept)    -0.851      0.0662        -12.9  8.71e-38
2 trip_mins       1.24       0.00897       138.   0
3 I(trip_mins^2) -0.00468    0.000259      -18.1  1.05e-72
4 I(trip_mins^3) -0.00000893 0.00000163     -5.48 4.24e- 8
```

More Complex Specifications: Polynomials

Visualising what's going on:



More Complex Specifications: Interactions

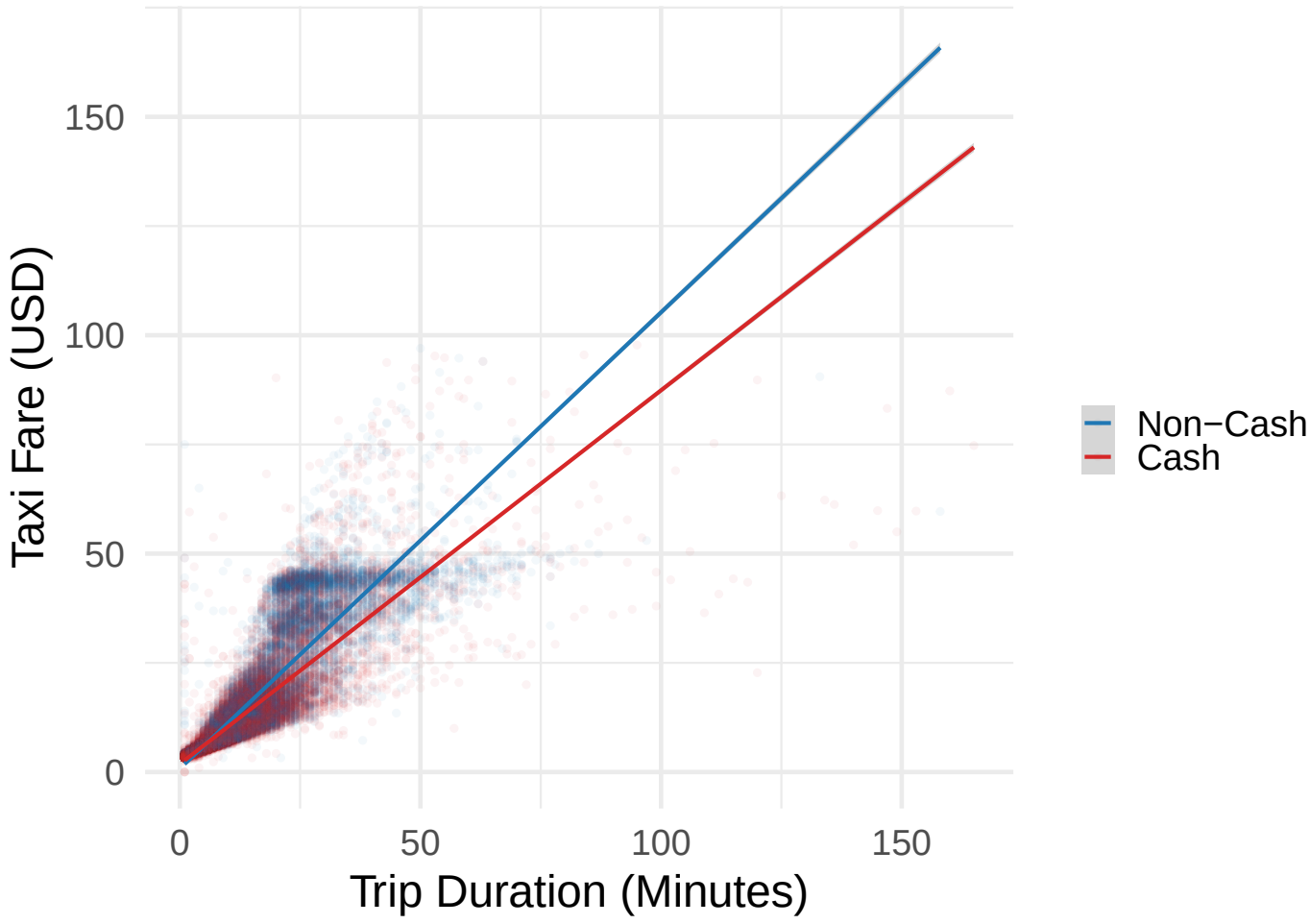
```
lm(fare ~ trip_mins * cash_dummy,
   data = taxi_trips) %>%
  summary() %>%
  broom::tidy()
```

```
# A tibble: 4 x 5
  term                 estimate std.error statistic   p.value
  <chr>                   <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)             0.789   0.0670       11.8 5.43e- 32
2 trip_mins               1.04    0.00383     273.  0
3 cash_dummy              1.03    0.0864       11.9 1.47e- 32
4 trip_mins:cash_dummy   -0.189   0.00535     -35.3 5.79e-269
```

Interpretation: Fare increases as trip_mins increases (marginal effect $= 1.04$), but less so (marginal effect $= 1.04 - 0.189 = 0.851$) when the trip is paid for in cash rather than alternatives.

More Complex Specifications: Interactions

Visually:



Inference

Classical Asymptotic Inference

Define a null hypothesis (H_0) and an alternative hypothesis (H_A).

H_0: Under this hypothesis, we believe that the true value of the parameter is zero ($\beta_1 = 0$)

H_A: Under this hypothesis, we believe that the true value of the parameter something other than zero ($\beta_1 \neq 0$)

Our goal is to ask, given the data we observe, how (un)likely is it that the H_0 is true?

Classical Asymptotic Inference

How do we do this?

1. Calculate a test-statistic of interest (often a t-statistic or F-statistic)
2. Assume an asymptotic distribution for the test-statistic (t-distribution or F-distribution) under the null hypothesis (drawing on the CLT)
3. Compare the observed test-statistic to the reference distribution
4. Calculate a p-value, which is the probability of observing a test-statistic as extreme as the one we observed, given that the null hypothesis is true.

Classical Asymptotic Inference

If the p-value is small, it is unlikely that our test-statistic is drawn from the null distribution.

Thus we might feel inclined to 'reject the null hypothesis' at some α level.

(α: our tolerance for false positives (Type I error). Typically set at 0.05, 0.01, or 0.001, but this is arbitrary.)

Classical Asymptotic Inference

To test linear regression coefficients we use the t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{H0}}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

where β_{H0} is the value of the parameter under the null hypothesis, $\hat{\beta}_1$ is the estimated coefficient, and $\text{se}(\hat{\beta}_1)$ is the standard error of the estimate.

Classical Asymptotic Inference

To test linear regression coefficients we use the t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{H0}}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

where β_{H0} is the value of the parameter under the null hypothesis, $\hat{\beta}_1$ is the estimated coefficient, and $\text{se}(\hat{\beta}_1)$ is the standard error of the estimate.

The standard error: Our estimate of the standard deviation of the sampling distribution of the estimator (CLT).

That is, a large standard error means the sampling distribution is more spread out – the estimator is higher variance.

Classical Asymptotic Inference

To test linear regression coefficients we use the t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{H0}}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

where β_{H0} is the value of the parameter under the null hypothesis, $\hat{\beta}_1$ is the estimated coefficient, and $\text{se}(\hat{\beta}_1)$ is the standard error of the estimate.

The standard error: Our estimate of the standard deviation of the sampling distribution of the estimator (CLT).

That is, a large standard error means the sampling distribution is more spread out – the estimator is higher variance.

So what is the t-statistic? Ratio of the magnitude of the estimate to the implicit variability of the estimate.

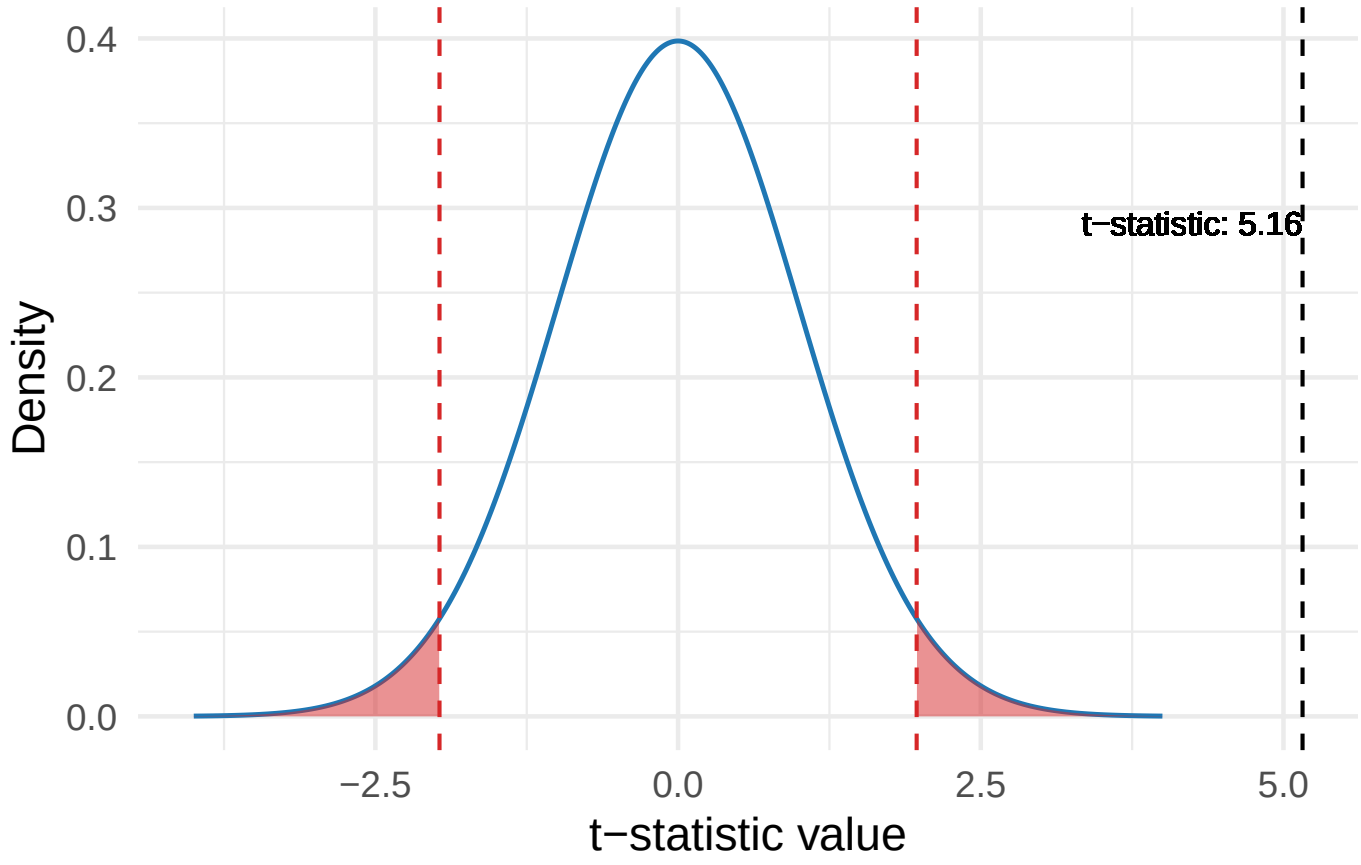
Classical Asymptotic Inference in Action

Calculating the t-statistic for our full regression:

```
# To make the t-stat more plausible, take a subsample:
taxi_sub <- taxi_trips %>%
  sample_frac(0.005)

# Fit a linear regression of tipped on trip_mins, extract t-stat
t_stat <- lm(fare ~ trip_mins + I(trip_mins^2) + I(trip_mins^3), data = taxi_su
  summary() %>%
  broom::tidy() %>%
  mutate(t_stat = estimate / std.error) %>%
  filter(term == "trip_mins") %>%
  select(t_stat)

t_stat
```

```
# A tibble: 1 x 1
  t_stat
   <dbl>
1   5.16
```

Classical Asymptotic Inference in Action



Confidence Intervals

We can use the standard error to construct a **confidence interval** for the parameter of interest, against the reference distribution (e.g. the t-distribution):

$$\hat{\beta}_1 \pm t_{crit} \cdot \text{se}(\hat{\beta}_1)$$

where t_{crit} is the critical value from the t-distribution at the desired confidence level (e.g. 95%, corresponding to $\alpha = 0.05$).

Interpretation: Over repeated samples, 95% of the 95% confidence intervals we construct will contain the true value of the parameter.

Heteroskedasticity-Robust Inference

In practice, homoskedasticity is a very implausible assumption.

In some cases, like the linear probability model, it is by definition violated.

Heteroskedasticity-Robust Inference

In practice, homoskedasticity is a very implausible assumption.

In some cases, like the linear probability model, it is by definition violated.

We can adjust for this by using **heteroskedasticity-robust standard errors**. Essentially, we allow each observation to have its own error variance, rather than assuming constant variance across observations.

There are various implementations in `R`:

- `{sandwich}::vcovHC()` used with `lm()`
- `{estimatr}::lm_robust()`
- `{fixest}::feols()`
- `{modelsummary}::modelsummary()`

General advice: Always estimate heteroskedasticity-robust standard errors.

Cluster-Robust Inference

Recall we had hierarchies in other datasets?

We might want to adjust for the fact that the residuals within each cluster are correlated.

We can do this by using **cluster-robust standard errors**.

Essentially, we allow our errors to be correlated within clusters, relaxing the i.i.d assumption.

Cluster-Robust Inference

Recall we had hierarchies in other datasets?

We might want to adjust for the fact that the residuals within each cluster are correlated.

We can do this by using **cluster-robust standard errors**.

Essentially, we allow our errors to be correlated within clusters, relaxing the i.i.d assumption.

The same packages support implementations in `R`:

- `{sandwich}::vcovHC()` used with `lm()`
- `{estimatr}::lm_robust()`
- `{fixest}::feols()`
- `{modelsummary}::modelsummary()`

The Bootstrap

The *nonparametric bootstrap* works as follows:

- B times:
  1. Draw a random sample X_b^* of size n from the original sample *with replacement*
  2. Estimate the chosen model
  3. Store the chosen test statistic (e.g. $\hat{\beta}_b^*$) from the fitted model.

With $\hat{\beta}^* = \beta_b^*, \hat{\ldots}, \beta_B^*$ in hand we can:

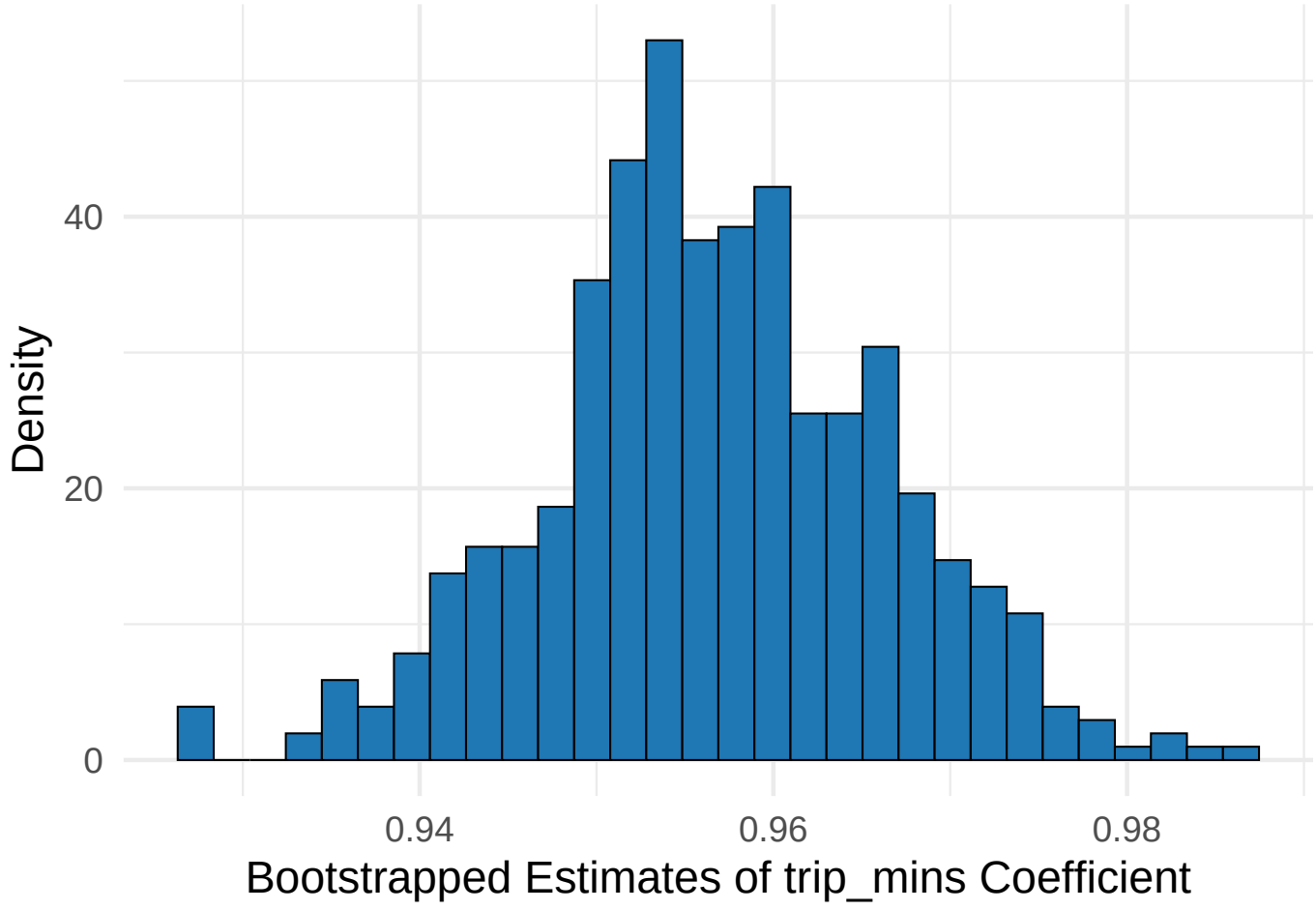
- Estimate the se of β as the standard deviation of $\hat{\beta}^*$
- Compute 95% confidence intervals using the 2.5th and 97.5th percentiles of $\hat{\beta}^*$

The Bootstrap in Action

```r
boot <- function(sims){
  # Create a vector to store the estimates
  estimates <- numeric(sims)

  for (i in 1:sims) {
    # Sample with replacement
    sample_data <- taxi_trips[sample(nrow(taxi_trips), replace = TRUE), ]
    # Fit the model
    model <- lm(fare ~ trip_mins, data = sample_data)
    # Store the estimate of the coefficient
    estimates[i] <- coef(model)["trip_mins"]
  }

  return(estimates)
}

boot_estimates <- boot(sims = 500)

sd(boot_estimates)
```

```
[1] 0.009748141
```

The Bootstrap in Action



This empirical distribution is our bootstrapped proxy of the sampling distribution of the estimator $\hat{\beta}_1$.