

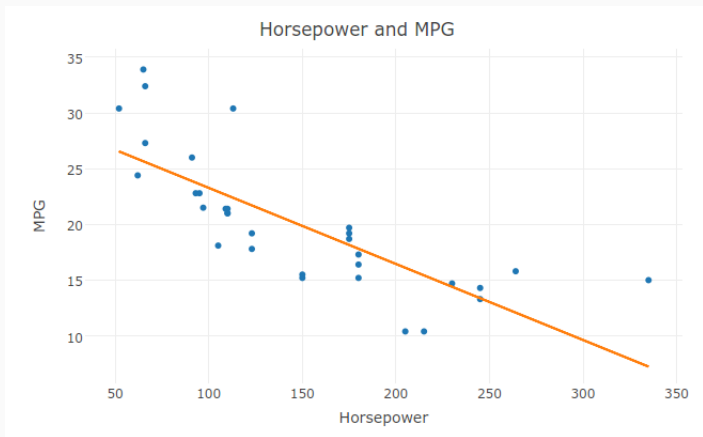
Econ 103 – Statistics for Economists

Chapter 9: Regression

Mallick Hossain

University of Pennsylvania

Recall: “Best Fitting” Line Through Cloud of Points



Recall: Regression as a Data Summary

Linear Model

$$\hat{y} = a + bx$$

Choose a, b to Minimize Sum of Squared Vertical Deviations

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

The Prediction

Predict a car's MPG given its horsepower

Recall: Regression as a Data Summary

Problem

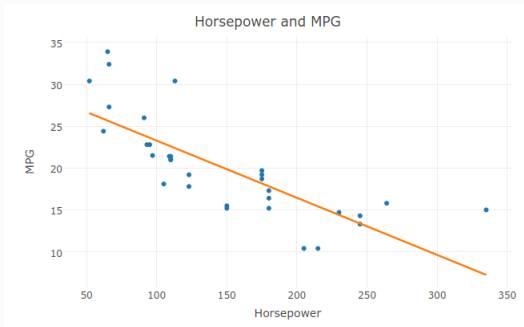
$$\min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

Solution

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = r \frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

Beyond Regression as a Data Summary



The estimated slope was about -0.07 hp/mpg and the estimated intercept was about 30 hp.

What if anything does this tell us about the relationship between horsepower and mpg *in the population*?

The Population Regression Model

How is Y (mpg) related to X (horsepower) in the population?

Assumption I: Linearity

The random variable Y is linearly related to X according to

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

β_0, β_1 are two unknown population parameters (constants).

Assumption II: Error Term ϵ

$E[\epsilon] = 0$, $Var(\epsilon) = \sigma^2$ and ϵ is independent of X . The error term ϵ measures the unpredictability of Y *after controlling for* X

Predictive Interpretation of Regression

Under Assumptions I and II

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

- “Best guess” of Y having observed $X = x$ is $\beta_0 + \beta_1 x$
- If $X = 0$, we predict $Y = \beta_0$
- If two people differ by one unit in X , we predict that they will differ by β_1 units in Y .

The only problem is, we don't know $\beta_0, \beta_1 \dots$

Estimating β_0, β_1

Suppose we observe an iid sample $(Y_1, X_1), \dots, (Y_n, X_n)$ from the population: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Then we can *estimate* β_0, β_1 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$$

Once we have estimators, we can think about sampling uncertainty...

Sampling Uncertainty: Pretend the Cars in Our Data is our Population

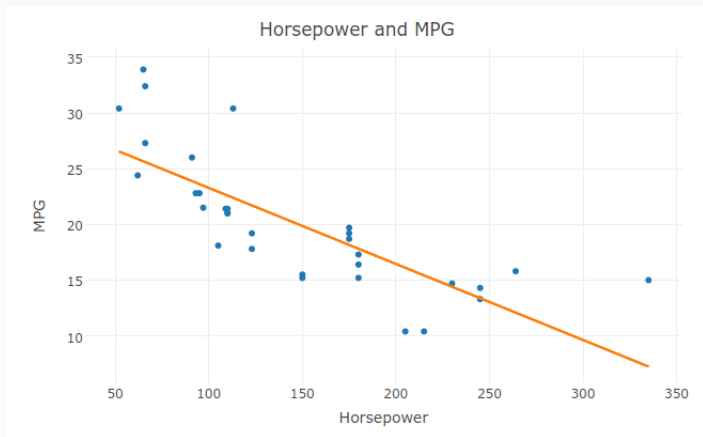
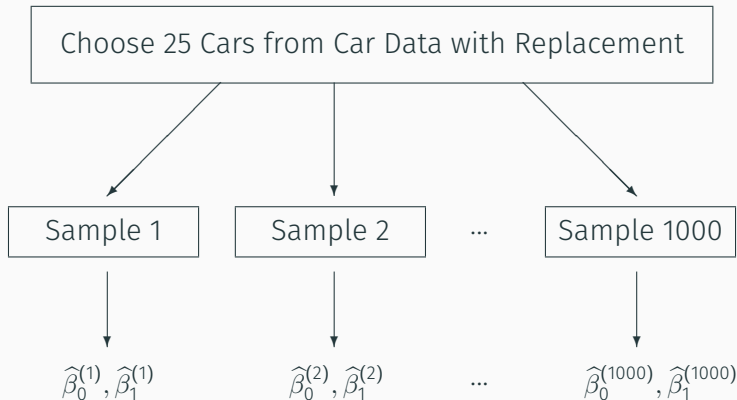


Figure 1: Estimated Slope = -0.07, Estimated Intercept = 30

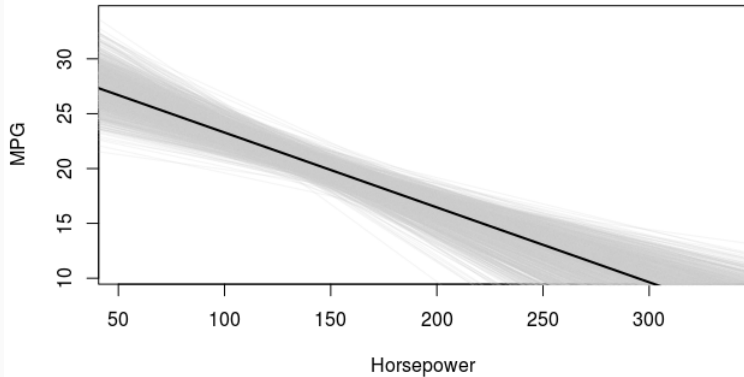
Sampling Distribution of Regression Coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$



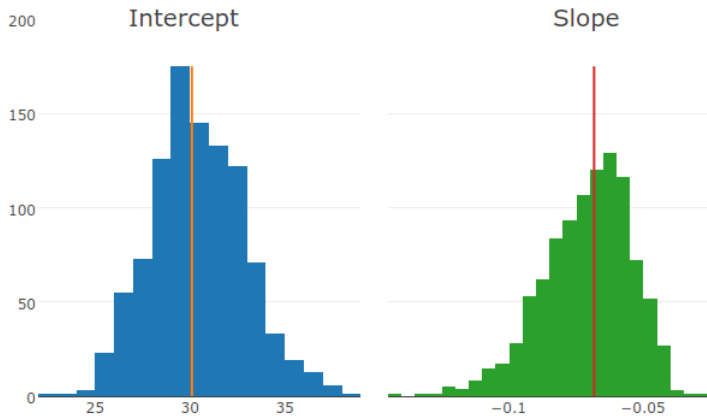
Repeat 1000 times → get 1000 different pairs of estimates

Sampling Distribution: long-run relative frequencies

1000 Replications, $n = 25$



Population: Intercept = 30, Slope = -0.07



Based on 1000 Replications, $n = 25$

Inference for Linear Regression

Central Limit Theorem

$$\frac{\hat{\beta} - \beta}{\widehat{SE}(\hat{\beta})} \approx N(0, 1)$$

How to calculate \widehat{SE} ?

- Complicated
 - Depends on variance of errors ϵ and all predictors in regression.
 - We'll look at a few simple examples
 - R does this calculation for us
- Requires assumptions about population errors ϵ_i
 - Simplest (and R default) is to assume $\epsilon_i \sim iid(0, \sigma^2)$
 - Weaker assumptions in Econ 104

Intuition for What Effects $SE(\hat{\beta}_1)$ for Simple Regression

$$SE(\hat{\beta}_1) \approx \frac{\sigma}{\sqrt{n}} \cdot \frac{1}{s_X}$$

- $\sigma = SD(\epsilon)$ – inherent variability of the Y , even after controlling for X
- n is the sample size
- s_X is the sampling variability of the X observations.

I treated the class as our population for the purposes of the simulation experiment but it makes more sense to think of the class as a sample from some population. We'll take this perspective now and think about various inferences we can draw from the cars data using regression.

$$\text{MPG} = \beta_0 + \epsilon$$

```
lm(formula = mpg ~ 1, data = mtcars)
      coef.est coef.se
(Intercept) 20.09      1.07
---
n = 32, k = 1
```


$$\text{MPG} = \beta_0 + \epsilon$$

```
lm(formula = mpg ~ 1, data = mtcars)
      coef.est coef.se
(Intercept) 20.09      1.07
---
n = 32, k = 1

> mean(mtcars$mpg)
[1] 20.09062
```

$$\text{MPG} = \beta_0 + \epsilon$$

```
lm(formula = mpg ~ 1, data = mtcars)
      coef.est coef.se
(Intercept) 20.09      1.07
---
n = 32, k = 1
```

```
> mean(mtcars$mpg)
[1] 20.09062
```

```
> sd(mtcars$mpg)/sqrt(length(mtcars$mpg))
[1] 1.065424
```

Dummy Variable (aka Binary Variable)

A predictor variable that takes on only two values: 0 or 1. Used to represent two categories, e.g. Automatic/Manual.

$$\text{MPG} = \beta_0 + \beta_1 \text{Manual} + \epsilon$$

```
lm(formula = mpg ~ am, data = mtcars)
      coef.est coef.se
(Intercept) 17.15    1.12
am           7.24    1.76
---
n = 32, k = 2
residual sd = 4.90, R-Squared = 0.36
```

$$\text{MPG} = \beta_0 + \beta_1 \text{Manual} + \epsilon$$

```
lm(formula = mpg ~ am, data = mtcars)
      coef.est coef.se
(Intercept) 17.15    1.12
am           7.24    1.76
---
n = 32, k = 2
residual sd = 4.90, R-Squared = 0.36

> mean(manual$mpg) - mean(automatic$mpg)
[1] 7.244939
```

$$\text{MPG} = \beta_0 + \beta_1 \text{Manual} + \epsilon$$

```
lm(formula = mpg ~ am, data = mtcars)
```

```
              coef.est coef.se  
(Intercept) 17.15      1.12  
am           7.24      1.76
```

```
---
```

```
n = 32, k = 2
```

```
residual sd = 4.90, R-Squared = 0.36
```

```
> mean(manual$mpg) - mean(automatic$mpg)
```

```
[1] 7.244939
```

```
> sqrt(var(manual$mpg)/length(manual$mpg) +  
        var(automatic$mpg)/length(automatic$mpg))
```

```
[1] 1.923202
```

$$\text{MPG} = \beta_0 + \beta_1 \text{Manual} + \epsilon$$

What is the ME for an approximate 95% confidence interval for the difference of population means of transmission: (automatic - manual)?

```
lm(formula = mpg ~ am, data = mtcars)
      coef.est coef.se
(Intercept) 17.15    1.12
am           7.24    1.76
---
n = 32, k = 2
residual sd = 4.90, R-Squared = 0.36
```

$$\text{MPG} = \beta_0 + \beta_1 \text{Horsepower} + \epsilon$$

```
lm(formula = mpg ~ hp, data = mtcars)
      coef.est coef.se
(Intercept) 30.10    1.63
hp          -0.07    0.01
---
n = 32, k = 2
residual sd = 3.86, R-Squared = 0.60
```


$$\text{MPG} = \beta_0 + \beta_1 \text{Horsepower} + \epsilon$$

What is the ME for an approximate 95% CI for β_1 ?

```
lm(formula = mpg ~ hp, data = mtcars)
      coef.est coef.se
(Intercept) 30.10    1.63
hp          -0.07    0.01
---
n = 32, k = 2
residual sd = 3.86, R-Squared = 0.60
```

Simple vs. Multiple Regression

Terminology

Y is the “outcome” and X is the “predictor.”

Simple Regression

One predictor variable: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Multiple Regression

More than one predictor variable:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i$$

- In both cases $\epsilon_1, \epsilon_2, \dots, \epsilon_n \sim \text{iid}(0, \sigma^2)$
- Multiple regression coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ calculated by minimizing sum of squared vertical deviations, but formula requires linear algebra so we won't cover it.

Interpreting Multiple Regression

Predictive Interpretation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i$$

β_j is the difference in Y that we would predict between two individuals who differed by one unit in predictor X_j *but who had the same values for the other X variables.*

What About an Example?

In a few minutes, we'll work through an extended example of multiple regression using real data.

Inference for Multiple Regression

In addition to estimating the coefficients $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ for us, R will calculate the corresponding standard errors. It turns out that

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\hat{\beta}_j)} \approx N(0, 1)$$

for *each* of the $\hat{\beta}_j$ by the CLT provided that the sample size is large.

$$\text{MPG} = \beta_0 + \beta_1 \text{Horsepower} + \epsilon$$

What are **residual sd** and **R-squared**?

```
lm(formula = mpg ~ hp, data = mtcars)
      coef.est coef.se
(Intercept) 30.10    1.63
hp          -0.07    0.01
---
n = 32, k = 2
residual sd = 3.86, R-Squared = 0.60
```

Fitted Values and Residuals

Fitted Value \hat{y}_i

Predicted y -value for a car i given its x -variables using estimated regression coefficients: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}$

Residual $\hat{\epsilon}_i$

Car i 's *vertical deviation* from regression line: $\hat{\epsilon}_i = y_i - \hat{y}_i$.

The residuals are *stand-ins* for the unobserved errors ϵ_i .

Residual Standard Deviation: $\hat{\sigma}$

- Idea: use residuals $\hat{\epsilon}_i$ to estimate σ

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n - k}}$$

- Measures avg. distance of y_i from regression line.
 - E.g. if Y is points scored on a test and $\hat{\sigma} = 16$, the regression predicts to an accuracy of about 16 points.
- Same units as Y (Exam practice: verify this)
- Denominator $(n - k) = (\# \text{ Datapoints} - \# \text{ of } X \text{ variables})$

Proportion of Variance Explained: R^2

$$R^2 \approx 1 - \frac{\widehat{\sigma^2}}{S_y^2}$$

- R^2 = proportion of $\text{Var}(Y)$ “explained” by the regression.
 - Higher value \implies greater proportion explained
- Unitless, between 0 and 1
- Generally harder to interpret than $\hat{\sigma}$, but...
- For simple linear regression $R^2 = (r_{xy})^2$ and this where its name comes from!

$$\text{MPG} = \beta_0 + \beta_1 \text{Horsepower} + \epsilon$$

```
lm(formula = mpg ~ hp, data = mtcars)
      coef.est coef.se
(Intercept) 30.10    1.63
hp          -0.07    0.01
---
n = 32, k = 2
residual sd = 3.86, R-Squared = 0.60
> cor(mtcars$mpg, mtcars$hp)^2
[1] 0.6024373
```

Which Gives Better Predictions: Transmission (a) or Horsepower (b)?

```
lm(formula = mpg ~ am, data = mtcars)
      coef.est coef.se
(Intercept) 17.15    1.12
am           7.24    1.76
---
n = 32, k = 2
residual sd = 4.90, R-Squared = 0.36
```

```
lm(formula = mpg ~ hp, data = mtcars)
      coef.est coef.se
(Intercept) 30.10    1.63
hp          -0.07    0.01
---
n = 32, k = 2
residual sd = 3.86, R-Squared = 0.60
```