# Simple Linear Regression

## Fundamental Techniques in Data Science with R

Utrecht University

Kyle M. Lang

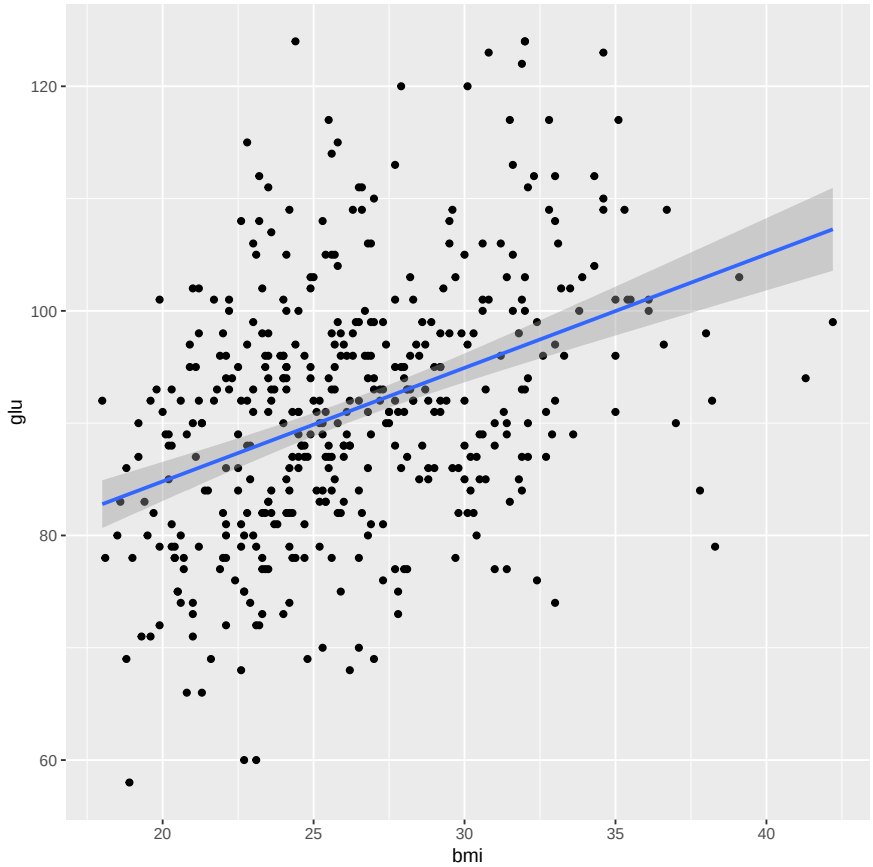
Department of Methodology & Statistics
Utrecht University

# Outline

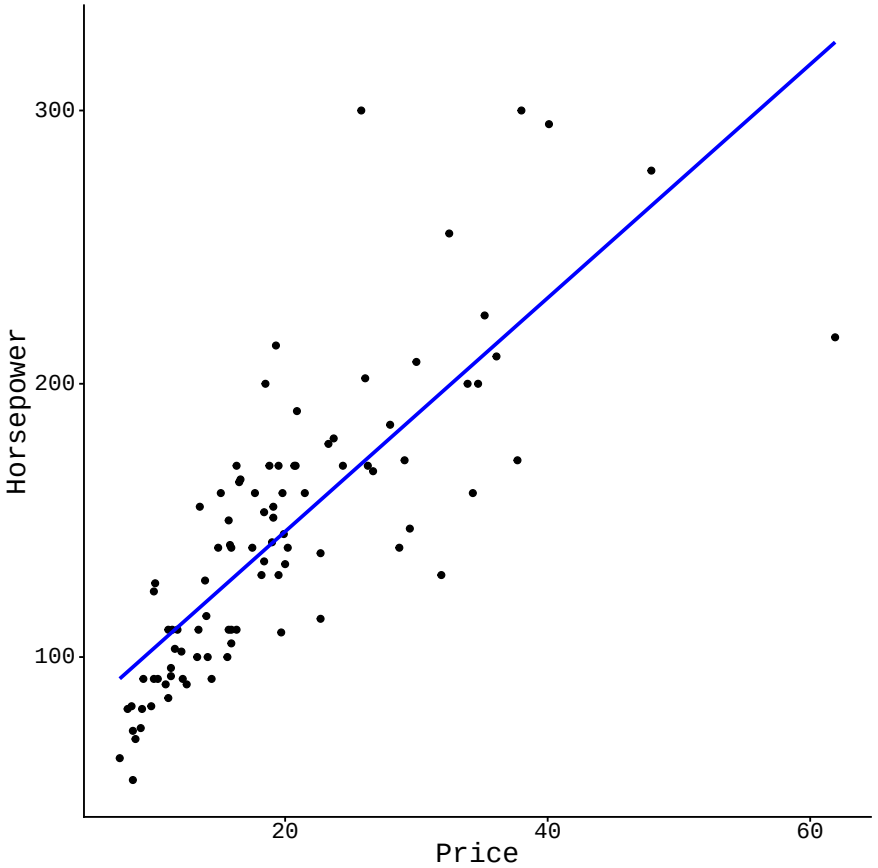
# Visualizations of Simple Linear Regression

Earlier, we used `geom_smooth()` to visualize the *best fit line* from a simple linear regression model.

```
library(ggplot2)

here::here("data", "diabetes.rds") |>
  readRDS() |>
  ggplot(aes(bmi, glu)) +
    geom_point() +
    geom_smooth(method = "lm")
```



# Visualizations of Simple Linear Regression



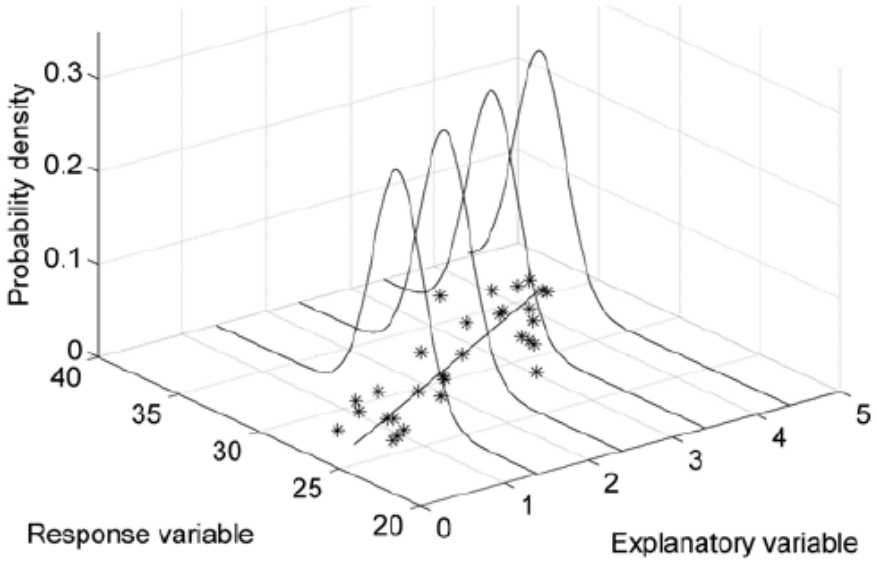




Image retrieved from:
`http://www.seaturtle.org/mtn/archives/mtn122/mtn122p1.shtml`

# Simple Linear Regression Equation

The best fit line is defined by a simple equation:

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

The above should look very familiar:

$$\begin{aligned} Y &= mX + b \\ &= \hat{\beta}_1 X + \hat{\beta}_0 \end{aligned}$$

$\hat{\beta}_0$ is the *intercept*.

- The $\hat{Y}$ value when $X = 0$.
- The expected value of $Y$ when $X = 0$.

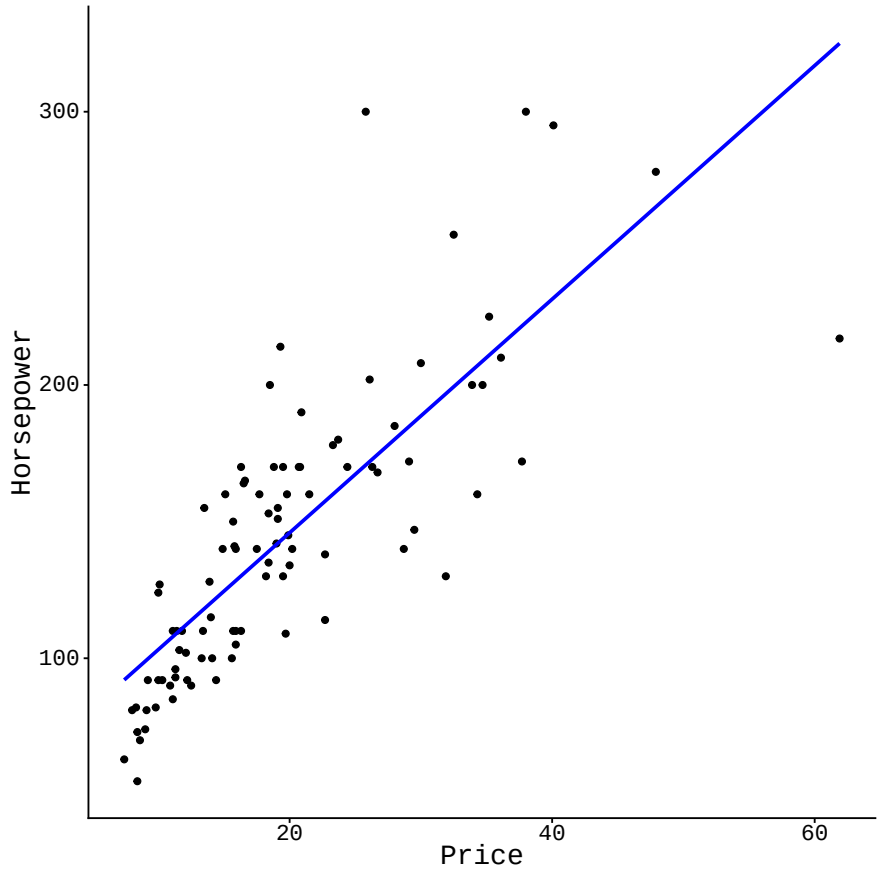
$\hat{\beta}_1$ is the *slope*.

- The change in $\hat{Y}$ for a unit change in $X$.
- The expected change in $Y$ for a unit change in $X$.

# Thinking about Error

The equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ only describes the best fit line.

- It does not fully quantify the relationship between $Y$ and $X$.



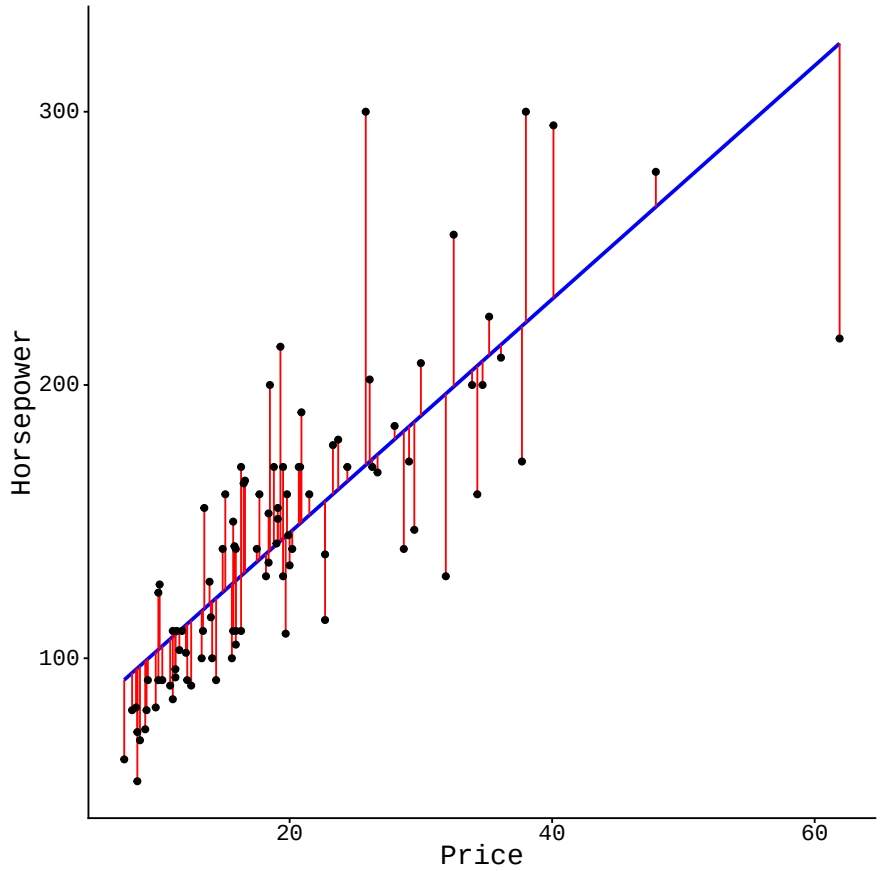
# Thinking about Error

The equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ only describes the best fit line.

- It does not fully quantify the relationship between $Y$ and $X$.

We still need to account for the estimation error.

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



# Estimating the Regression Coefficients

The purpose of regression analysis is to use a sample of $N$ observed $\{Y_n, X_n\}$ pairs to find the best fit line defined by $\hat{\beta}_0$ and $\hat{\beta}_1$.

- The most popular method of finding the best fit line involves minimizing the sum of the squared residuals.
- $RSS = \sum_{n=1}^{N} \hat{\varepsilon}_n^2$

# Residuals as the Basis of Estimation

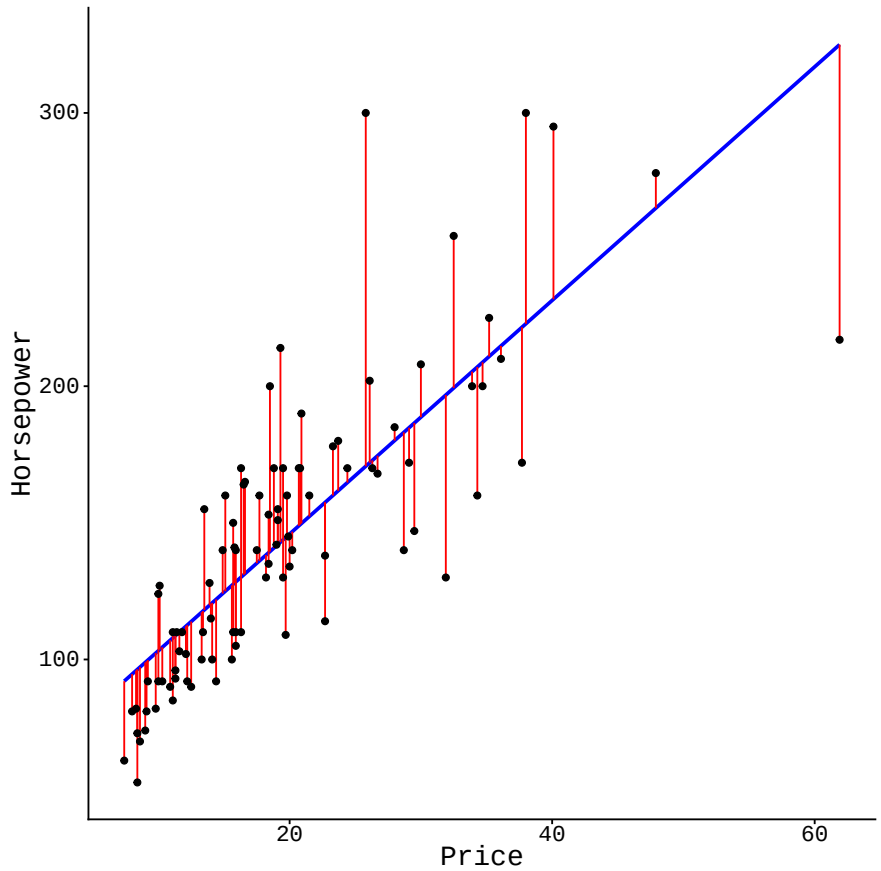
The $\hat{\varepsilon}_n$ are defined in terms of deviations between each observed $Y_n$ value and the corresponding $\hat{Y}_n$.

$$\hat{\varepsilon}_n = Y_n - \hat{Y}_n = Y_n - \left(\hat{\beta}_0 + \hat{\beta}_1 X_n\right)$$

Each $\hat{\varepsilon}_n$ is squared before summing to remove negative values.

$$RSS = \sum_{n=1}^{N} \hat{\varepsilon}_n^2 = \sum_{n=1}^{N} \left(Y_n - \hat{Y}_n\right)^2$$

$$= \sum_{n=1}^{N} \left(Y_n - \hat{\beta}_0 - \hat{\beta}_1 X_n\right)^2$$



# Least Squares Example

Estimate the least squares coefficients for our example data:

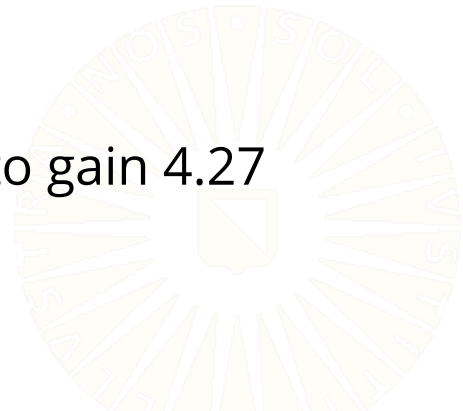
```
data(Cars93, package = "MASS")

out1 <- lm(Horsepower ~ Price, data = Cars93)
coef(out1)

(Intercept)       Price
  60.447578    4.273796
```

The estimated intercept is $\hat{\beta}_0 = 60.45$.

- A free car is expected to have 60.45 horsepower.

The estimated slope is: $\hat{\beta}_1 = 4.27$.

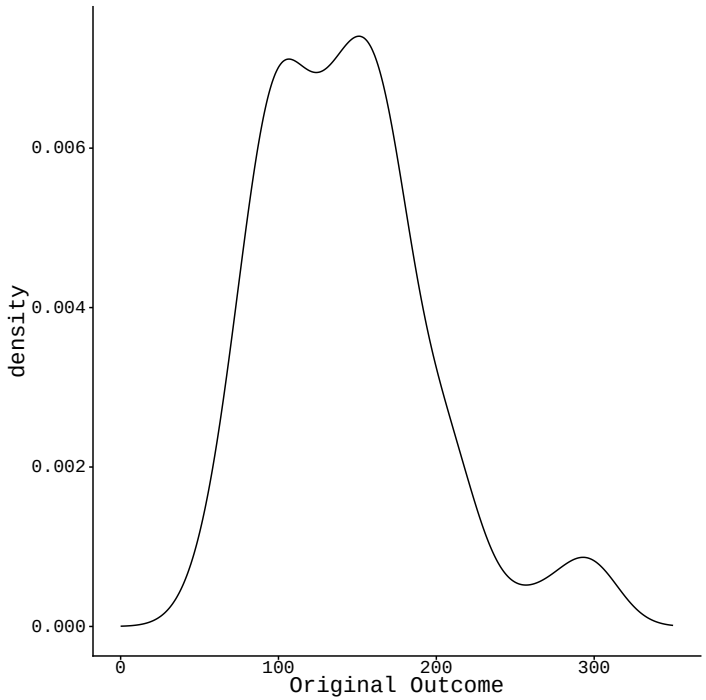
- For every additional $1000 in price, a car is expected to gain 4.27 horsepower.

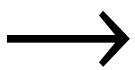


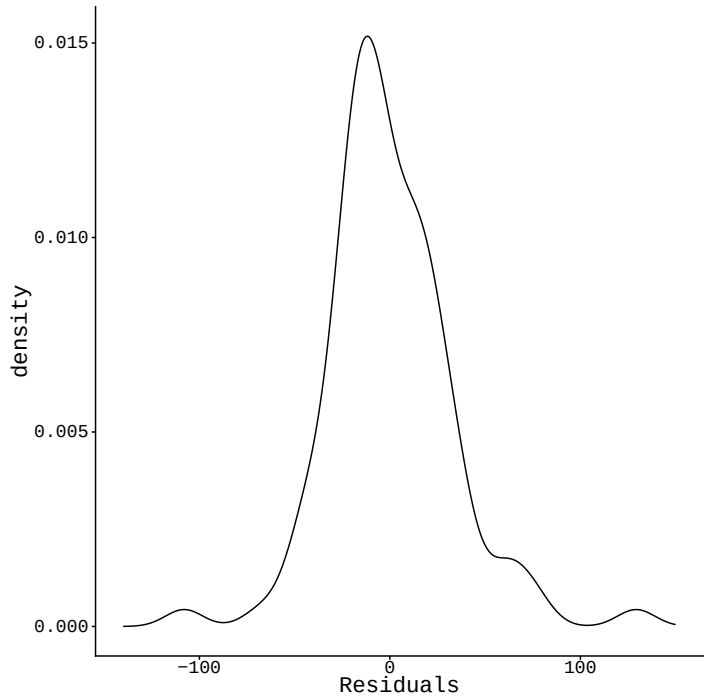
# Model Fit

We may also want to know how well our model explains the outcome.

- Our model explains some proportion of the outcome's variability.
- The residual variance $\hat{\sigma}^2 = \mathbf{Var}(\hat{\varepsilon})$ will be less than $\mathbf{Var}(Y)$.



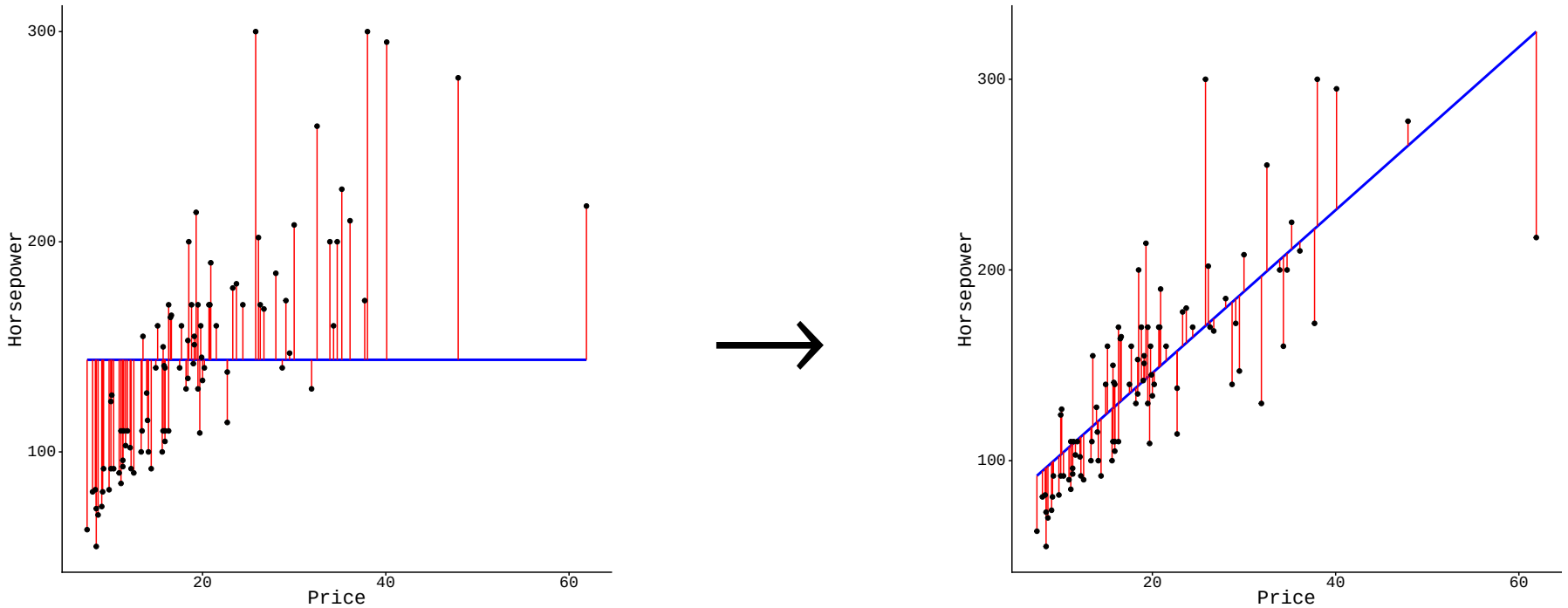





# Model Fit

We may also want to know how well our model explains the outcome.

- Our model explains some proportion of the outcome's variability.
- The residual variance $\hat{\sigma}^2 = \mathrm{Var}(\hat{\varepsilon})$ will be less than $\mathrm{Var}(Y)$.



# Model Fit

We quantify the proportion of the outcome's variance that is explained by our model using the $R^2$ statistic:

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

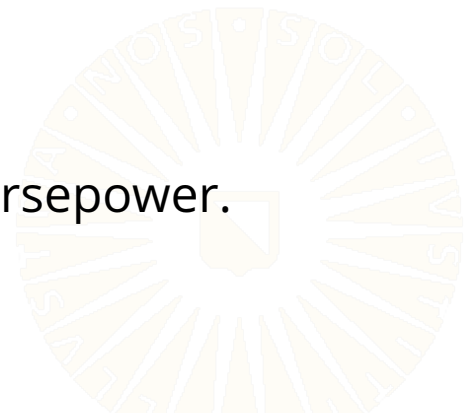
where

$$TSS = \sum_{n=1}^{N} \left(Y_n - \bar{Y}\right)^2 = \text{Var}(Y) \times (N - 1)$$

For our example problem, we get:

$$R^2 = 1 - \frac{95573}{252363} \approx 0.62$$

Indicating that car price explains 62% of the variability in horsepower.



# Model Fit for Prediction

When assessing predictive performance, we will most often use the *mean squared error* (MSE) as our criterion.

$$
\begin{aligned}
MSE &= \frac{1}{N}\sum_{n=1}^{N}\left(Y_n - \hat{Y}_n\right)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(Y_n - \hat{\beta}_0 - \sum_{p=1}^{P}\hat{\beta}_p X_{np}\right)^2 \\
&= \frac{RSS}{N}
\end{aligned}
$$

For our example problem, we get:

$$
MSE = \frac{95573}{93} \approx 1027.67
$$

# Interpreting MSE

The MSE quantifies the average squared prediction error.

- Taking the square root improves interpretation.

$$RMSE = \sqrt{MSE}$$

The RMSE estimates the magnitude of the expected prediction error.

- For our example problem, we get:

$$RMSE = \sqrt{\frac{95573}{93}} \approx 32.06$$

- When using price as the only predictor of horsepower, we expect prediction errors with magnitudes of 32.06 horsepower.

# Information Criteria

We can use *information criteria* to quickly compare *non-nested* models while accounting for model complexity.

- Akaike's Information Criterion (AIC)

$$AIC = 2K - 2\hat{\ell}(\theta|X)$$

- Bayesian Information Criterion (BIC)

$$BIC = K\ln(N) - 2\hat{\ell}(\theta|X)$$

# Information Criteria

We can use *information criteria* to quickly compare *non-nested* models while accounting for model complexity.

- Akaike's Information Criterion (AIC)

$$AIC = 2K - 2\hat{\ell}(\theta|X)$$

- Bayesian Information Criterion (BIC)

$$BIC = K\ln(N) - 2\hat{\ell}(\theta|X)$$

Information criteria balance two competing forces.

- The optimized loglikelihood quantifies fit to the data.
- The penalty term corrects for model complexity.

# Information Criteria

For our example, we get the following estimates of AIC and BIC:

$$\begin{aligned} AIC &= 2(3) - 2(-454.44) \\ &= 914.88 \end{aligned}$$

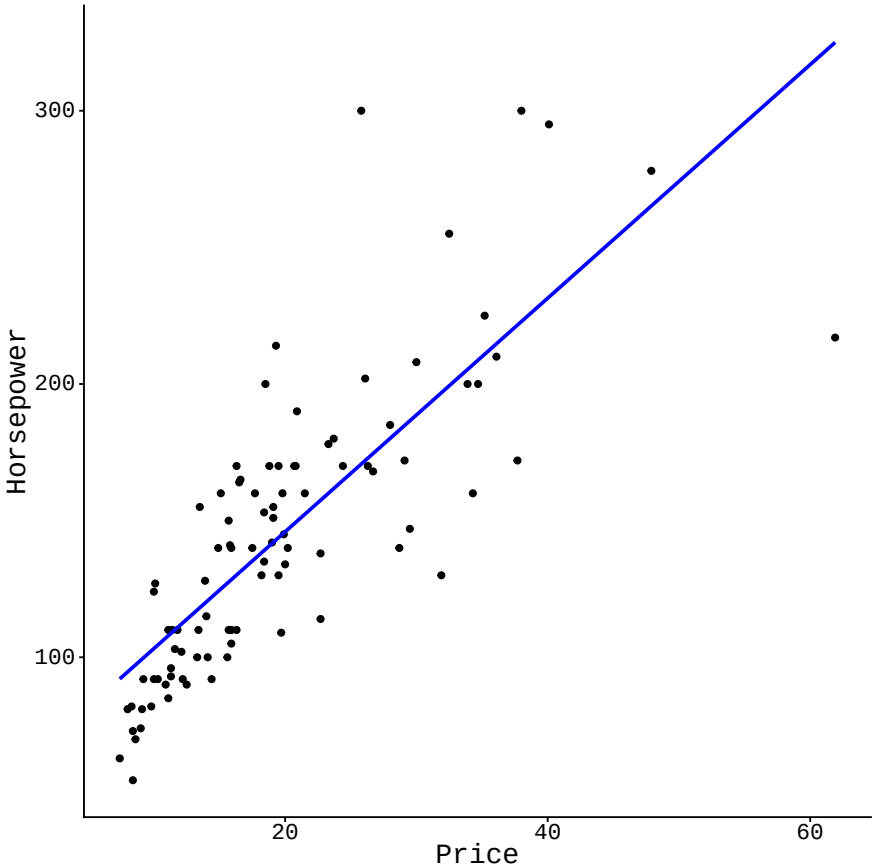
$$\begin{aligned} BIC &= 3\ln(93) - 2(-454.44) \\ &= 922.48 \end{aligned}$$

To compute the AIC/BIC from a fitted `lm()` object in R:

```
AIC(out1)

[1] 914.8821

BIC(out1)

[1] 922.4799
```

# Linear Regression as a Statistical Model

Consider the implications of these two ways of visualizing a simple linear regression analysis.



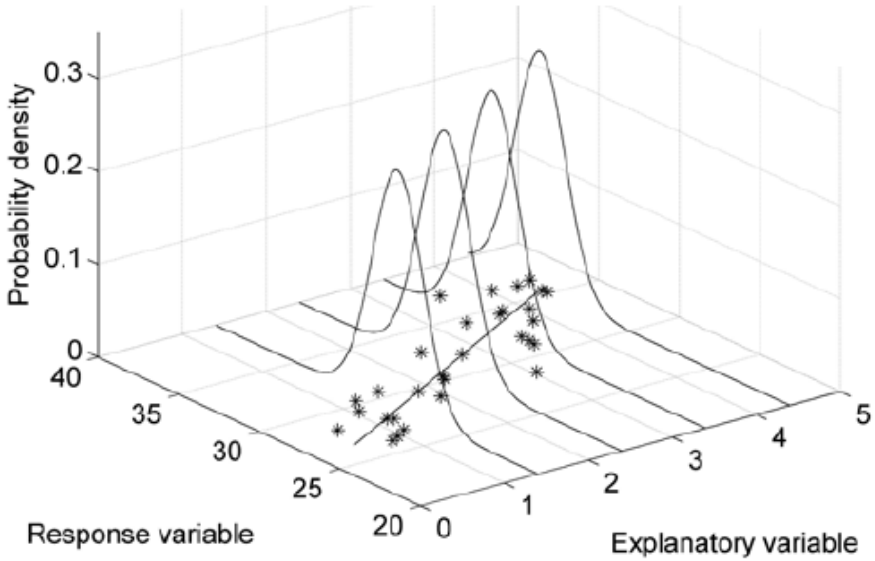




Image retrieved from:
`http://www.seaturtle.org/mtn/archives/mtn122/mtn122p1.shtml`

# Linear Regression as a Statistical Model

When we adopt a modeling perspective, we formulate our linear regression equation as a probability model.

Partition the systematic and random components:

$$Y = \eta + \varepsilon$$
$$\eta = \beta_0 + \beta_1 X$$
$$\varepsilon \sim \mathrm{N}(0, \sigma^2)$$

Or, more succinctly:

$$Y \sim \mathrm{N}(\eta, \sigma^2)$$
$$\eta = \beta_0 + \beta_1 X$$