Simple Linear Regression

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

The “Regression” Problem

Simple Linear Regression

Model Fit

Regression Problem

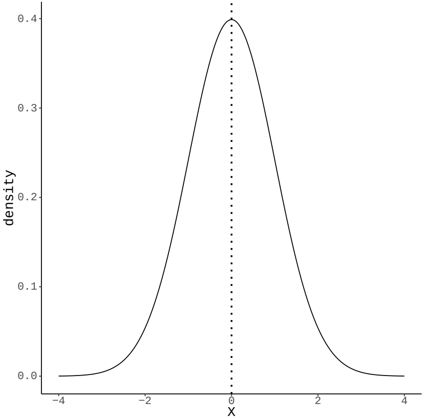
Some of the most ubiquitous and useful statistical models are *regression models*.

- *Regression* problems (as opposed to *classification* problems) involve modeling a quantitative response.
- The regression problem begins with a random outcome variable, Y.
- We hypothesize that the mean of Y is dependent on some set of fixed covariates, $\mathbf{X}$.

Flavors of Probability Distribution

The distributions with which you're probably most familiar imply a constant mean.

- Each observation is expected to have the same value of Y, regardless of their individual characteristics.
- This type of distribution is called "marginal" or "unconditional."



Flavors of Probability Distribution

The distributions we consider in regression problems have *conditional means*.

- The value of Y that we expect for each observation is defined by the observations' individual characteristics.
- This type of distribution is called "conditional."

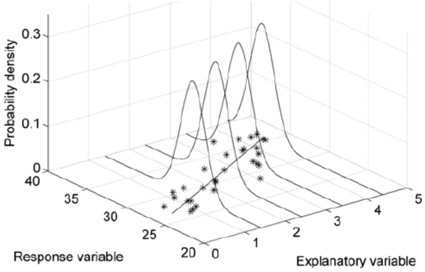
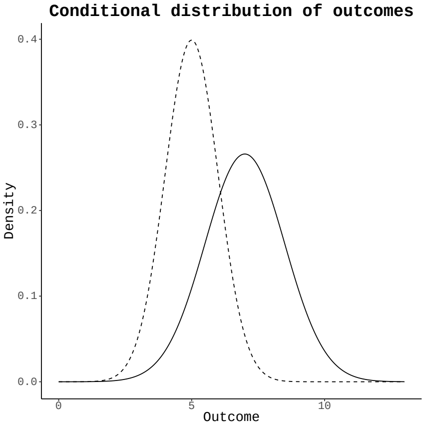


Image retrieved from:
`http://www.seaturtle.org/mtn/archives/mtn122/mtn122p1.shtml`

Flavors of Probability Distribution

Even a simple comparison of means implies a conditional distribution.

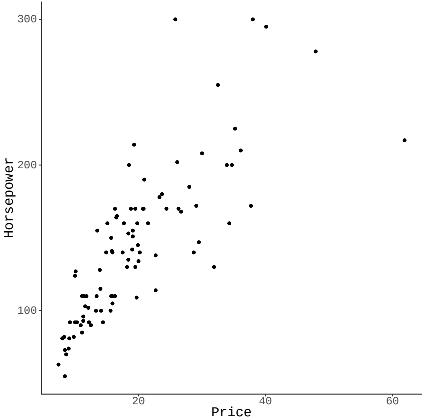
- The solid curve corresponds to outcome values for one group.
- The dashed curve represents outcomes from the other group.



Projecting a Distribution onto the Plane

In practice, we only interact with the X-Y plane of the previous 3D figure.

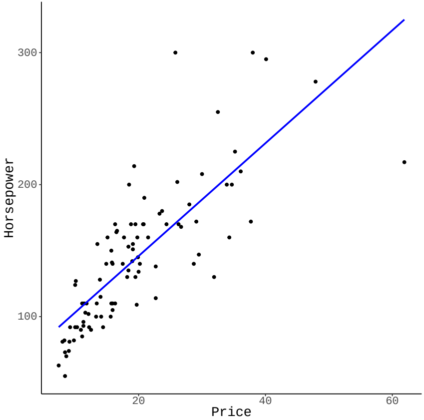
- On the Y-axis, we plot our outcome variable
- The X-axis represents the predictor variable upon which we condition the mean of Y.



Modeling the X–Y Relationship in the Plane

We want to explain the relationship between Y and X by finding the line that traverses the scatterplot as "closely" as possible to each point.

- This is the "best fit line".
- For any given value of X the corresponding point on the best fit line is our best guess for the value of Y, given the model.



Simple Linear Regression

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 $X = 0$.

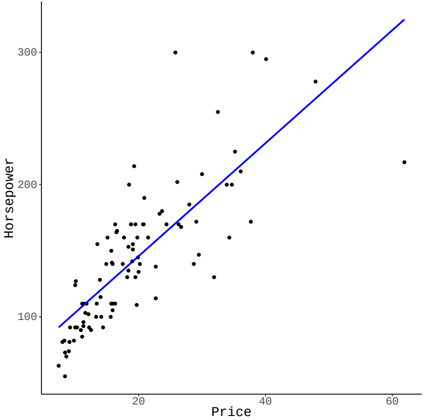
$\hat{\beta}_1$ is the *slope*.

- 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.



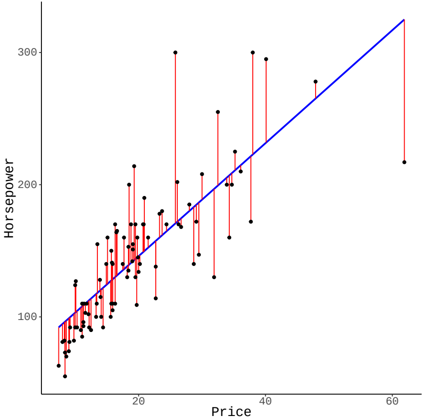
Thinking about Error

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.
- $RSS = \sum_{n=1}^{N} \hat{\varepsilon}_n^2$

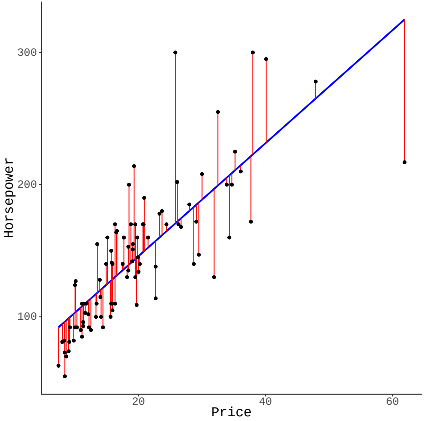
Residuals as the Basis of Estimation

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)
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-Based Prediction

In the social and behavioral sciences, regression modeling is often focused on inference about estimated model parameters.

- The association between the price of a car and its power.
- We model the system and scrutinize $\hat{\beta}_1$ to make inferences about the association between price and power.

Model-Based Prediction

In the social and behavioral sciences, regression modeling is often focused on inference about estimated model parameters.

- The association between the price of a car and its power.
- We model the system and scrutinize $\hat{\beta}_1$ to make inferences about the association between price and power.

In data science applications, we're often more interested in predicting the outcome for new observations.

- After we estimate $\hat{\beta}_0$ and $\hat{\beta}_1$, we can plug in new predictor data and get a predicted outcome value for any new case.
- In our example, these predictions represent the projected horsepower ratings of cars with prices given by the new X_{price} values.

Inference vs. Prediction

When doing statistical inference, we focus on how certain variables relate to the outcome.

- Do men have higher job-satisfaction than women?
- Does increased spending on advertising correlate with more sales?
- Is there a relationship between the number of liquor stores in a neighborhood and the amount of crime?

Inference vs. Prediction

When doing statistical inference, we focus on how certain variables relate to the outcome.

- Do men have higher job-satisfaction than women?
- Does increased spending on advertising correlate with more sales?
- Is there a relationship between the number of liquor stores in a neighborhood and the amount of crime?

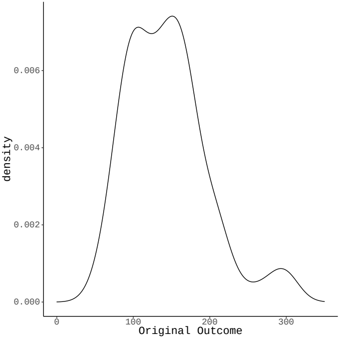
When doing prediction, we want to build a tool that can accurately guess future values.

- Will it rain tomorrow?
- Will this investment turn a profit within one year?
- Will increasing the number of contact hours improve grades?

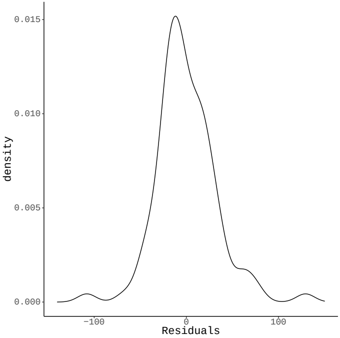
Model Fit

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 $\mathrm{Var}(Y)$.



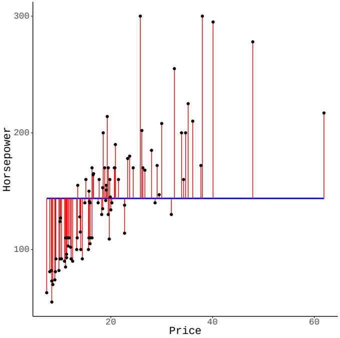
$\longrightarrow$



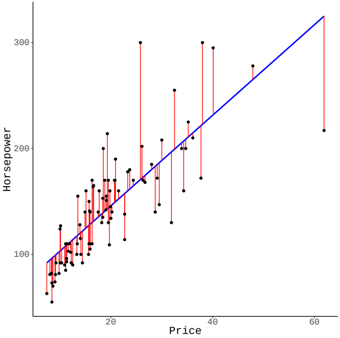
Model Fit

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 $\hat{\sigma}^2 = \text{Var}(\hat{\varepsilon})$ will be less than $\text{Var}(Y)$.



$\longrightarrow$



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.