

# Linear Regression Basics

Utrecht University Winter School: Regression in R



**Utrecht  
University**

Kyle M. Lang

Department of Methodology & Statistics  
Utrecht University

2022-02-02

# Outline

---

Regression Problem

Simple Linear Regression

- Model Estimation
- Model Fit

Multiple Linear Regression

- Model Comparison



# 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 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."

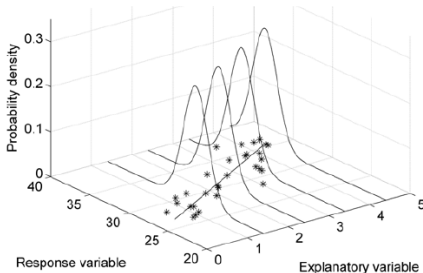
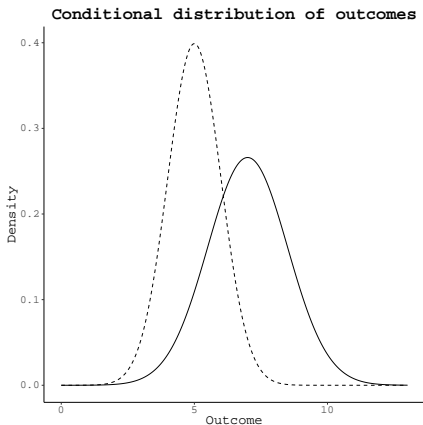


Image retrieved from:  
<http://www.seaturtle.org/mtn/archives/mtn122/mtn122p1.shtml>

# 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.

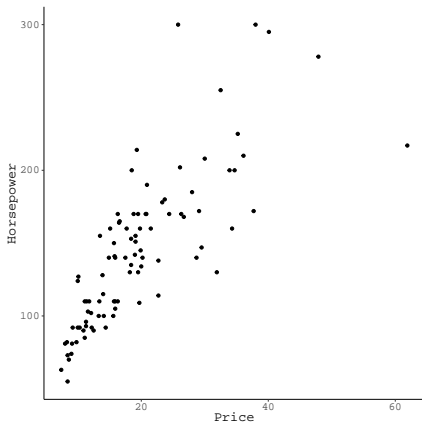


# SIMPLE LINEAR REGRESSION

# Projecting a Distribution onto the Plane

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

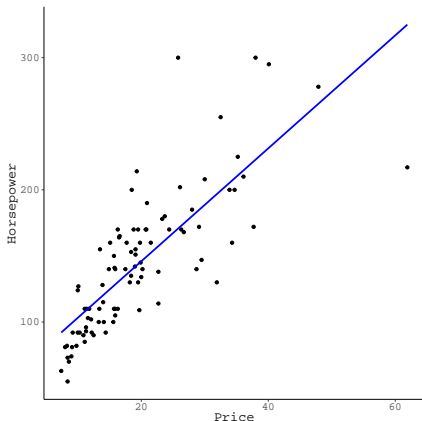
- On the Y-axis, we plot our outcome variable
- The X-axis represents the predictor variable upon which we condition the mean of  $Y$ .



# 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.

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





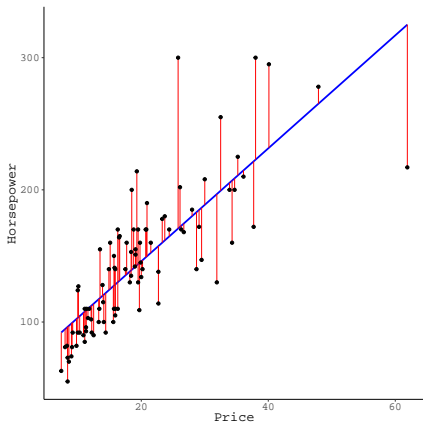
# Modeling the X-Y Relationship in the Plane

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.

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

To fully model the relation between  $Y$  and  $X$ , we still need to account for the estimation error.

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



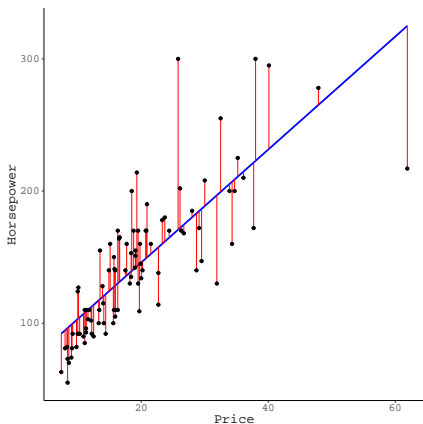
# Residuals as the Basis of Estimation

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

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

Each  $\hat{\varepsilon}_n$  is squared before summing to produce a quadratic objective function.

$$\begin{aligned} RSS &= \sum_{n=1}^N \hat{\varepsilon}_n^2 = \sum_{n=1}^N (Y_n - \hat{Y}_n)^2 \\ &= \sum_{n=1}^N (Y_n - \hat{\beta}_0 - \hat{\beta}_1 X_n)^2 \end{aligned}$$



# 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:  $\hat{\beta}_1 = 4.27$ .

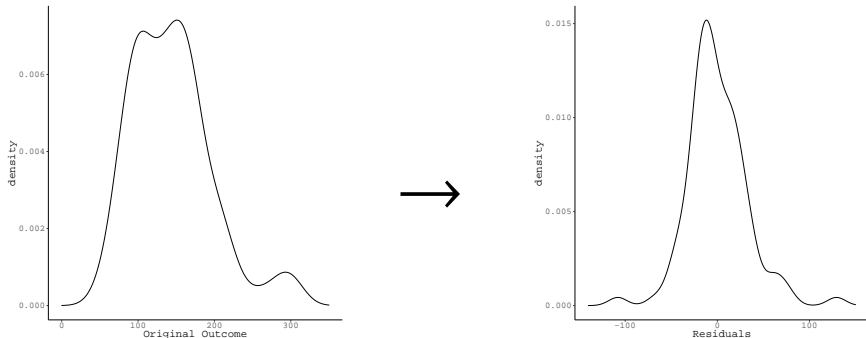
- For every additional \$1000 in price, a car is expected to gain 4.27 horsepower.



# Model Fit

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

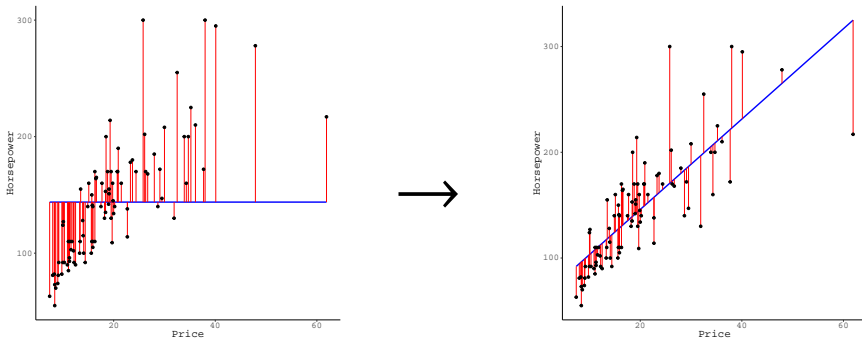
- Our model explains some proportion of the outcome's variability.
- The residual variance  $\hat{\sigma}^2 = \text{Var}(\hat{\varepsilon})$  will be less than  $\text{Var}(Y)$ .



# 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  $\text{Var}(Y)$ .



# 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

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

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.

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

For our example problem, we get:

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



# Interpreting MSE

---

The MSE quantifies the average squared prediction error.

- Taking the square root improves interpretation.

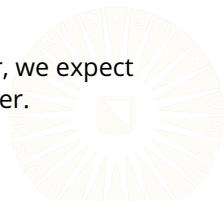
$$RMSE = \sqrt{MSE}$$

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)
```

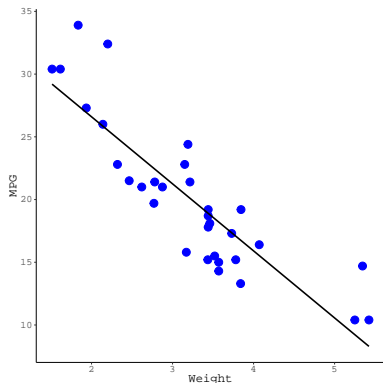
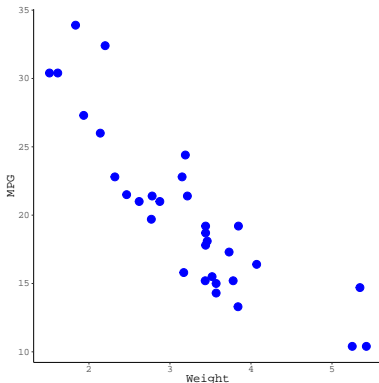
```
[1] 922.4799
```

# MULTIPLE LINEAR REGRESSION

# Graphical Representations of Regression Models

A regression of two variables can be represented on a 2D scatterplot.

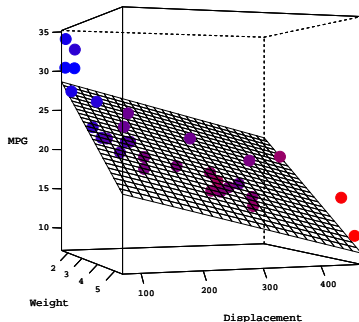
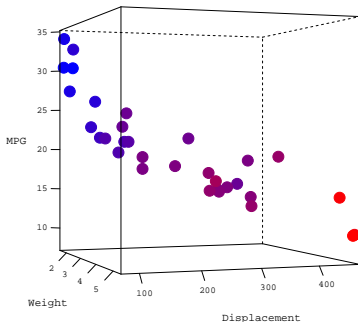
- Simple linear regression implies a 1D line in 2D space.



# Graphical Representations of Regression Models

Adding an additional predictor leads to a 3D point cloud.

- A regression model with two IVs implies a 2D plane in 3D space.



# Partial Effects

---

In MLR, we want to examine the *partial effects* of the predictors.

- What is the effect of a predictor after controlling for some other set of variables?

This approach is crucial to controlling confounds and adequately modeling real-world phenomena.

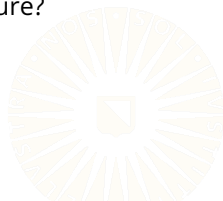


# Example

---

```
## Read in the 'diabetes' dataset:  
dataDir <- "../.../data/"  
dDat    <- readRDS(paste0(dataDir, "diabetes.rds"))  
  
## Simple regression with which we're familiar:  
out1 <- lm(bp ~ age, data = dDat)
```

Asking: What is the effect of age on average blood pressure?





# Example

---

```
partSummary(out1, -1)
```

Residuals:

Min	1Q	Median	3Q	Max
-31.188	-8.897	-1.209	8.612	39.952

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	77.47605	2.38132	32.535	< 2e-16
age	0.35391	0.04739	7.469	4.39e-13

Residual standard error: 13.04 on 440 degrees of freedom

Multiple R-squared: 0.1125, Adjusted R-squared: 0.1105

F-statistic: 55.78 on 1 and 440 DF, p-value: 4.393e-13

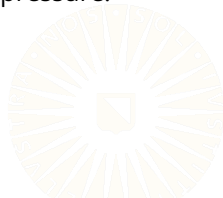
# Example

---

```
## Add in another predictor:  
out2 <- lm(bp ~ age + bmi, data = dDat)
```

Asking: What is the effect of BMI on average blood pressure, *after controlling for age*?

- We're partialing age out of the effect of BMI on blood pressure.



# Example

```
partSummary(out2, -1)
```

Residuals:

Min	1Q	Median	3Q	Max
-29.287	-8.198	-0.178	8.413	41.026

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	52.24654	3.83168	13.635	< 2e-16
age	0.28651	0.04504	6.362	5.02e-10
bmi	1.08053	0.13363	8.086	6.06e-15

Residual standard error: 12.18 on 439 degrees of freedom

Multiple R-squared: 0.2276, Adjusted R-squared: 0.224

F-statistic: 64.66 on 2 and 439 DF, p-value: < 2.2e-16

# Multiple $R^2$

---

How much variation in blood pressure is explained by the two models?

- Check the  $R^2$  values.

```
## Extract  $R^2$  values:  
r2.1 <- summary(out1)$r.squared  
r2.2 <- summary(out2)$r.squared  
  
r2.1  
[1] 0.1125117  
  
r2.2  
[1] 0.2275606
```

# F-Statistic

---

How do we know if the  $R^2$  values are significantly greater than zero?

- We use the F-statistic to test  $H_0 : R^2 = 0$  vs.  $H_1 : R^2 > 0$ .

```
f1 <- summary(out1)$fstatistic
```

```
f1
```

value	numdf	dendf
55.78116	1.00000	440.00000

```
pf(q = f1[1], df1 = f1[2], df2 = f1[3], lower.tail = FALSE)
```

value
4.392569e-13

# F-Statistic

---

```
f2 <- summary(out2)$fstatistic
```

```
f2
```

value	numdf	dendf
64.6647	2.0000	439.0000

```
pf(f2[1], f2[2], f2[3], lower.tail = FALSE)
```

value
2.433518e-25

# Comparing Models

---

How do we quantify the additional variation explained by BMI, above and beyond age?

- Compute the  $\Delta R^2$

```
## Compute change in R^2:
```

```
r2.2 - r2.1
```

```
[1] 0.115049
```

# Significance Testing

How do we know if  $\Delta R^2$  represents a significantly greater degree of explained variation?

- Use an  $F$ -test for  $H_0 : \Delta R^2 = 0$  vs.  $H_1 : \Delta R^2 > 0$

```
## Is that increase significantly greater than zero?  
anova(out1, out2)
```

Analysis of Variance Table

Model 1: bp ~ age

Model 2: bp ~ age + bmi

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	440	74873				
2	439	65167	1	9706.1	65.386	6.057e-15 ***

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



# Model Comparison

---

We can also compare models based on their prediction errors.

- For OLS regression, we usually compare MSE values.

```
mse1 <- MSE(y_pred = predict(out1), y_true = dDat$bp)
mse2 <- MSE(y_pred = predict(out2), y_true = dDat$bp)
```

```
mse1
```

```
[1] 169.3963
```

```
mse2
```

```
[1] 147.4367
```

In this case, the MSE for the model with *BMI* included is smaller.

- We should prefer the the larger model.

# Model Comparison

---

Finally, we can compare models based on information criteria.

```
AIC(out1, out2)
```

	df	AIC
out1	3	3528.792
out2	4	3469.424

```
BIC(out1, out2)
```

	df	BIC
out1	3	3541.066
out2	4	3485.789

In this case, both the AIC and the BIC for the model with *BMI* included are smaller.

- We should prefer the the larger model.