Review of Linear Regression

Fundamental Techniques in Data Science with R

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Regression Problem

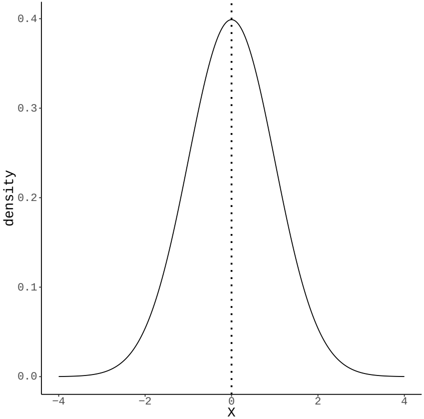
Some of the most ubiquitous and useful statistical models are *regression models*.

- *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."

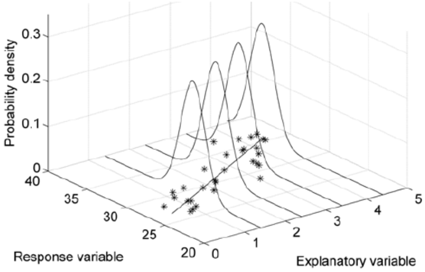
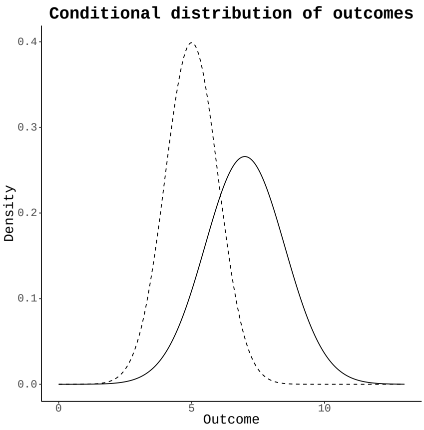


Image retrieved from:
http://www.seaturtle.org/mtn/archives/mtn122/mtn122p1.shtml

Flavors of Probability Distribution

Even a simple comparison of means implies a conditional distribution.

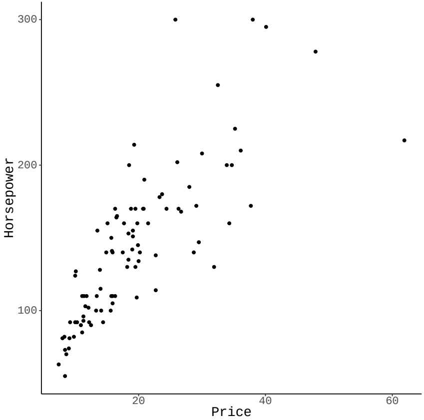
- The solid curve corresponds to outcome values for one group.
- The dashed curve represents outcomes from the other group.



Projecting a Distribution onto the Plane

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

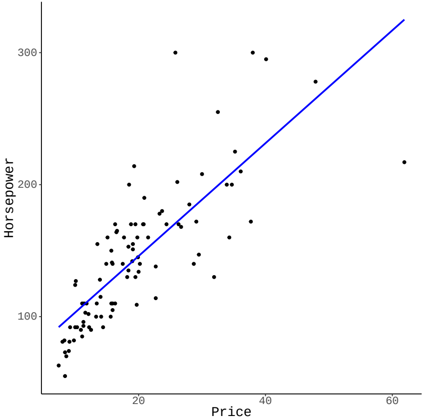
- On the Y-axis, we plot our outcome variable
- The X-axis represents the predictor variable upon which we condition the mean of Y.



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

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 $X = 0$.

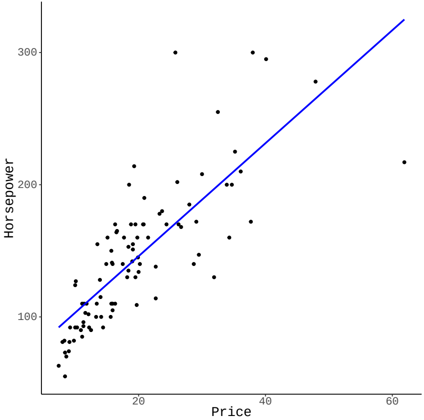
$\hat{\beta}_1$ is the *slope*.

- 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.



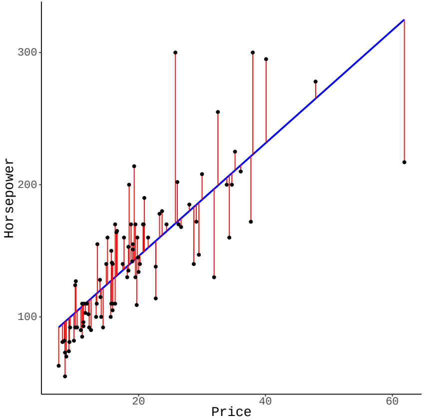
Thinking about Error

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.

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



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

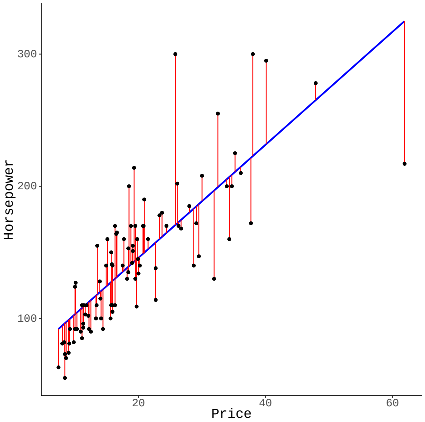
Residuals as the Basis of Estimation

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

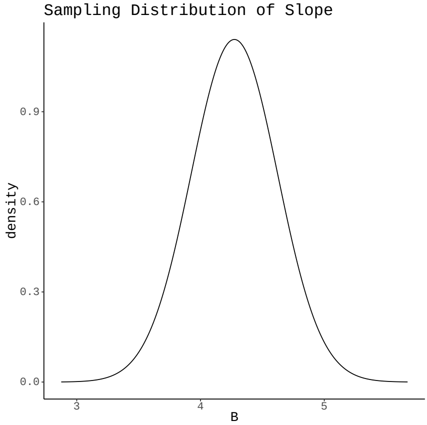
- For every additional $1000 in price, a car is expected to gain 4.27 horsepower.

Inference for Linear Models

Sampling Distribution

A sampling distribution is simply the probability distribution of a parameter.

- The *population* is defined by an infinite sequence of repeated estimations.
    - The sampling distribution quantifies the possible values of the statistic over infinite repeated sampling.
- The area of a region under the curve represents the probability of observing a *statistic* within the corresponding interval.



Some intuition: `http://onlinestatbook.com/stat_sim/sampling_dist/`

Test Statistics

If we want to "test" a slope coefficient, $\hat{\beta}$, we need some point of comparison.
- The *null-hypothesized* value of the slope, $H_0 : \beta = \tilde{\beta}$. - What value would the slope take if our hypothesis were false?

Our hypothesis test is actually a test for the size of the difference: $\hat{\beta} - \beta$
- We define a *test statistic*, t, to quantify the size of this effect accounting for the precision with which we've estimated $\hat{\beta}$.

We can construct the test statistic for $\hat{\beta}$ as follows:

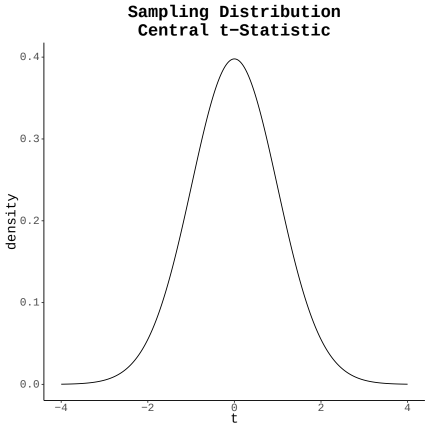
$$t = \frac{\hat{\beta} - \tilde{\beta}}{SE\left(\hat{\beta}\right)} \overset{\tilde{\beta}=0}{\Longrightarrow} t = \frac{\hat{\beta} - 0}{SE\left(\hat{\beta}\right)} = \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)}$$

For the slope in our example, we get a test statistic of:

$$t = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)} = \frac{4.27}{0.35} = 12.2$$

Sampling Distribution of Test Statistic

The t-statistic also has a sampling distribution that quantifies the possible t-values we could get if we repeatedly drew samples from the variables' distributions and re-computed a t-statistic each time.

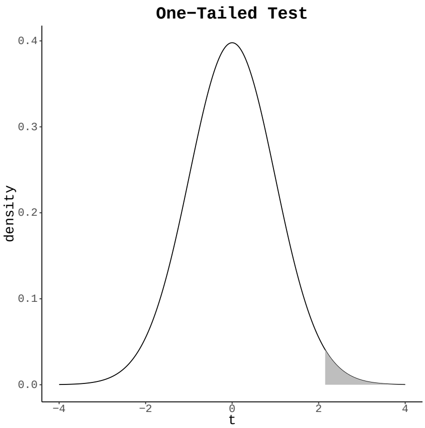


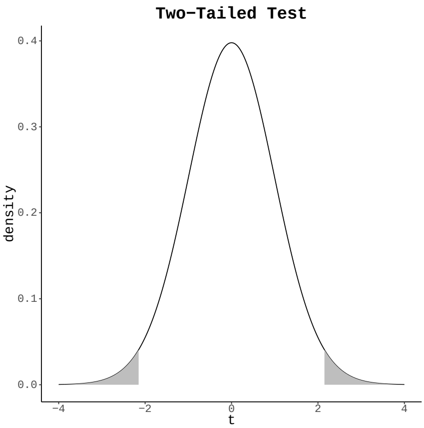
P-Values

Once we compute our estimated test statistic, $\hat{t}$, we compare it to the appropriate null-hypothesized sampling distribution.

- By calculating the area in the null distribution that exceeds our estimated test statistic, we can compute the probability of observing the given test statistic, or one more extreme, if the null hypothesis were true.
  - In other words, we can compute the probability of having sampled the data we observed, or more unusual data, from a population wherein there is no true difference between $\hat{\beta}$ and $\tilde{\beta}$.
- This value is the infamous *p-value*.

P-Values


One−Tailed Test
0.4
0.3
0.2
0.1
0.0
density
−4
−2
0
2
4
t


Two−Tailed Test
0.4
0.3
0.2
0.1
0.0
density
−4
−2
0
2
4
t

Interpreting P-Values

Consider the one-tailed test for our estimated test-statistic of $\hat{t} = 2.15$ that produces a p-value of $p = 0.017$.

- We *cannot* say that there is a 0.017 probability that the true mean difference is greater than zero.
- We *cannot* say that there is a 0.017 probability that the alternative hypothesis is true.
- We *cannot* say that there is a 0.017 probability that the null hypothesis is false.
- We *cannot* say that there is a 0.017 probability that the observed result is due to chance alone.
- We *cannot* say that there is a 0.017 probability of replicating the observed effect in future studies.

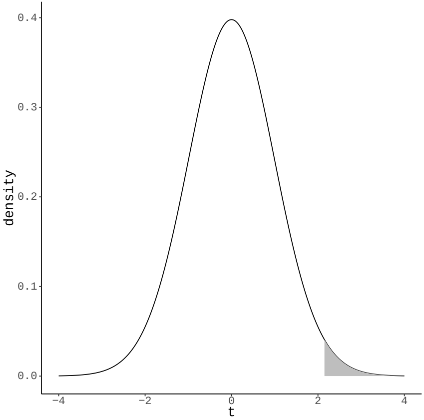
Interpreting P-Values

The p-value tells us $P(t \geq \hat{t} | H_0)$

- What we really want to know is $P(H_0 | t \geq \hat{t})$.

All that we *can* say is that there is a 0.017 probability of observing a test statistic at least as large as $\hat{t}$, if the null hypothesis is true.

- Our test uses the same logic as *proof by contradiction*.



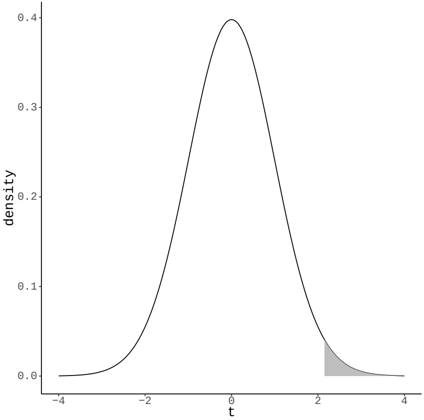
Interpreting P-Values

Note that $P(t \geq \hat{t}|H_0) \neq P(t = \hat{t}|H_0)$

- We *cannot* say that there is a 0.017 probability of observing $\hat{t}$, if the null hypothesis is true.

The probability of observing any individual point on a continuous distribution is exactly zero.

- $P(t = \hat{t}|H_0) = 0$



Inferential Tools: Confidence Intervals

We can also construct confidence intervals by:

$$CI = \hat{\beta} \pm t_{crit} \times SE\left(\hat{\beta}\right)$$

For our example slope, we get a 95% CI of:

$$CI_{95} = 4.27 \pm 1.99 \times 0.35 = [3.57; 4.97]$$

Which suggests that we can be 95% certain that the true value of β_1 is somewhere between 3.57 and 4.97.

- We are *95% certain* in the sense that if we repeat this analysis an infinite number of times, 95% of the CIs that we calculate will surround the true value of β_1.

Interpreting Confidence Intervals

Say we estimate a regression slope of $\hat{\beta}_1 = 0.5$ with an associated 95% confidence interval of $CI = [0.25; 0.75]$.

Interpreting Confidence Intervals

Say we estimate a regression slope of $\hat{\beta}_1 = 0.5$ with an associated 95% confidence interval of $CI = [0.25; 0.75]$.

- We *cannot* say that there is 95% chance that the true value of β_1 is between 0.25 and 0.75.
- We *cannot* say that the true value of β_1 is between 0.25 and 0.75, with probability 0.95.

Interpreting Confidence Intervals

Say we estimate a regression slope of $\hat{\beta}_1 = 0.5$ with an associated 95% confidence interval of $CI = [0.25; 0.75]$.

- We *cannot* say that there is 95% chance that the true value of β_1 is between 0.25 and 0.75.
- We *cannot* say that the true value of β_1 is between 0.25 and 0.75, with probability 0.95.

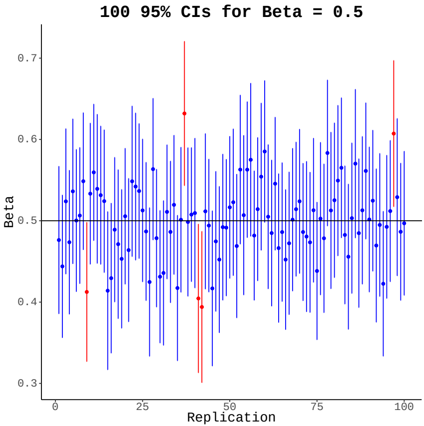
The true value of β_1 is fixed; it's a single quantity.

- β_1 is either in our estimated interval or it is not; there is no uncertainty.
- The probability that β_1 is within our estimated interval is either exactly 1 or exactly 0.

Interpreting Confidence Intervals

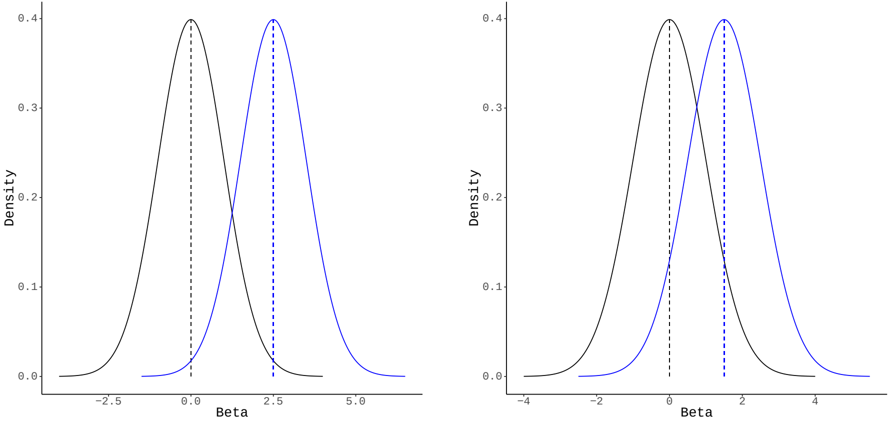
We don't talk about 95% probabilities when interpreting CIs; instead, we talk about 95% confidence.

- If we collected a new sample—of the same size—re-estimated our model, and re-computed the 95% CI for $\hat{\beta}_1$, we would get a different interval.
- Repeating this process an infinite number of times would give us a distribution of CIs.
- 95% of those CIs would surround the true value of β_1.



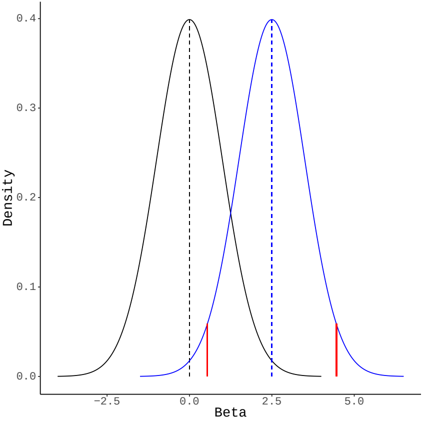
Inference with Confidence Intervals

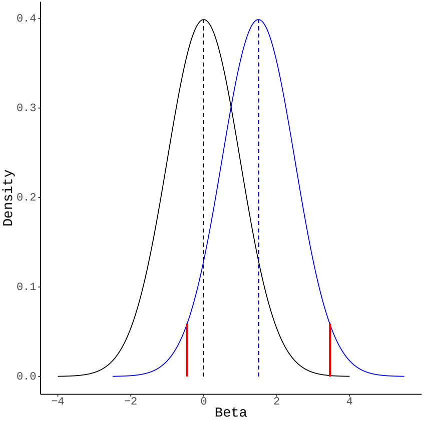
In terms of sampling distributions, our inferential task is to say something about how distinct the null and alternative distributions are.



Inference with Confidence Intervals

CIs give us a plausible range for the population value of β, so we can use CIs to support inference.





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 (or classification), we want to build a tool that can accurately guess future values.

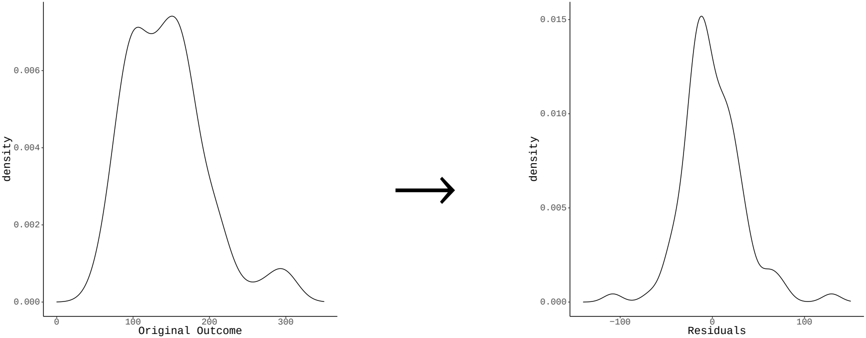
- Will it rain tomorrow?
- How much will a company earn from investing in a certain research profile?
- What is a patients risk of heart disease based on their medical history and test results?

Model Fit

Model Fit

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

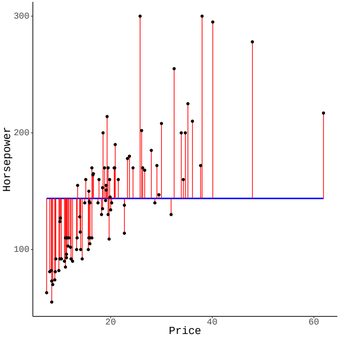
- Our model explains some proportion of the outcome's variability.
- The residual variance $\hat{\sigma}^2 = \text{Var}(\hat{\varepsilon})$ will be less than $\text{Var}(Y)$.

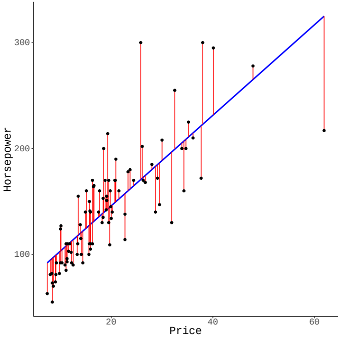


Model Fit

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

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

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.

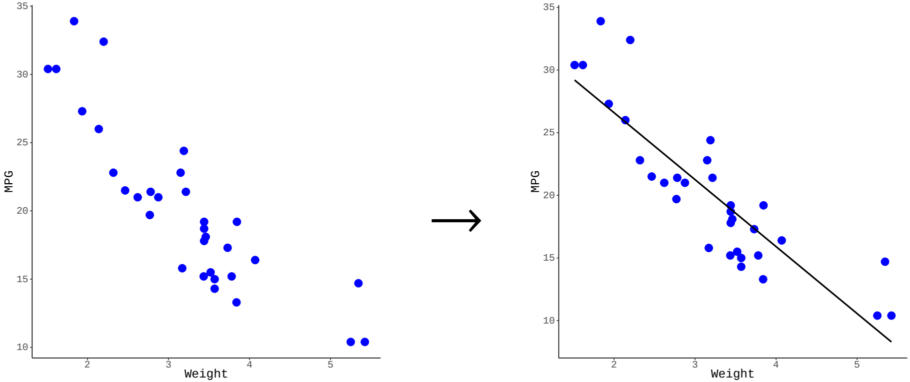
AIC

Multiple Linear Regression

Graphical Representations of Regression Models

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

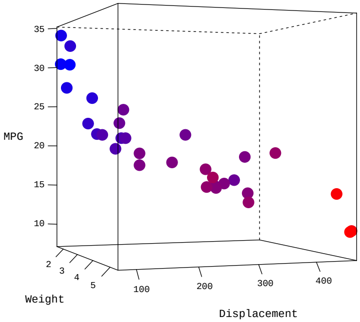
- Simple linear regression implies a 1D line in 2D space.



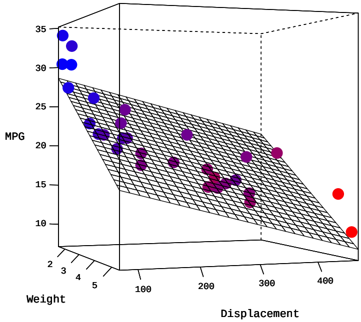
Graphical Representations of Regression Models

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

```r
## Read in the 'diabetes' dataset:
dDat <- readRDS("../data/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

```r
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

```r
## 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.

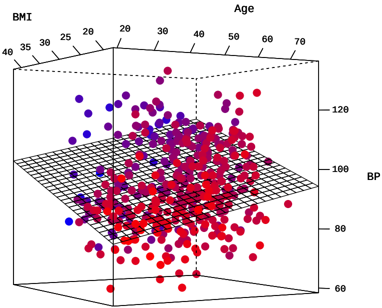
Example

```
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
```

Interpretation

- The expected average blood pressure for an unborn patient with a negligible extent is 52.25.
- For each year older, average blood pressure is expected to increase by 0.29 points, after controlling for BMI.
- For each additional point of BMI, average blood pressure is expected to increase by 1.08 points, after controlling for age.



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 ΔR^2

```
## Compute change in R^2:
r2.2 - r2.1

## [1] 0.115049
```

Significance Testing

How do we know if ΔR^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.