Introduction to Linear Modeling

Fundamental Techniques in Data Science with R

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Simple Linear Regression
- Model Fit

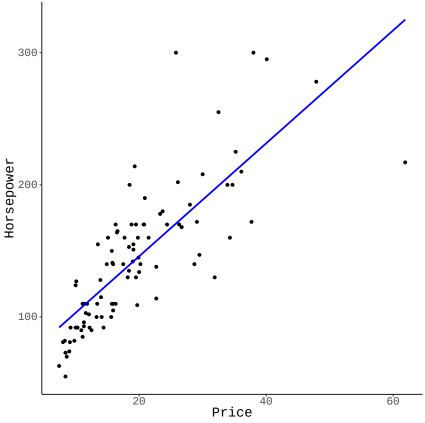
Multiple Linear Regression
- Model Comparison

Categorical Predictors
- Significance Testing for Dummy Codes

Model-Based Prediction
- Interval Estimates for Prediction

Moderation
- Categorical Moderators

Visualizations of Simple Linear Regression



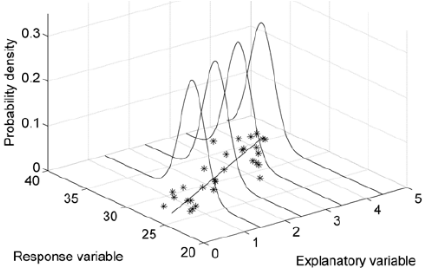


Image retrieved from:
http://www.seaturtle.org/mtn/archives/mtn122/mtn122p1.shtml

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 $X = 0$.

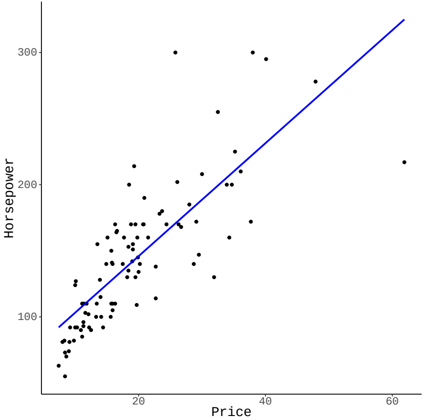
$\hat{\beta}_1$ is the *slope*.

- 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.



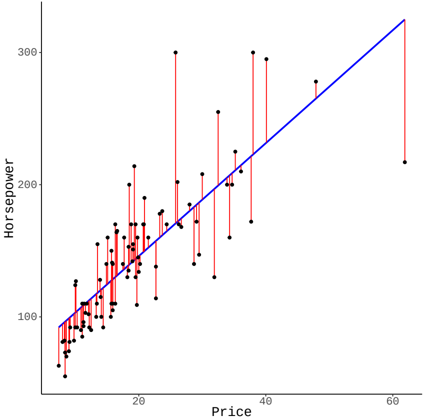
Thinking about Error

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

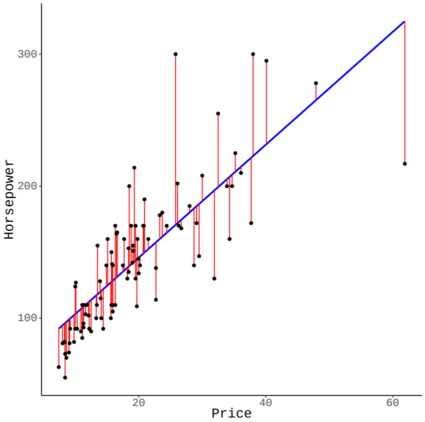
Residuals as the Basis of Estimation

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.

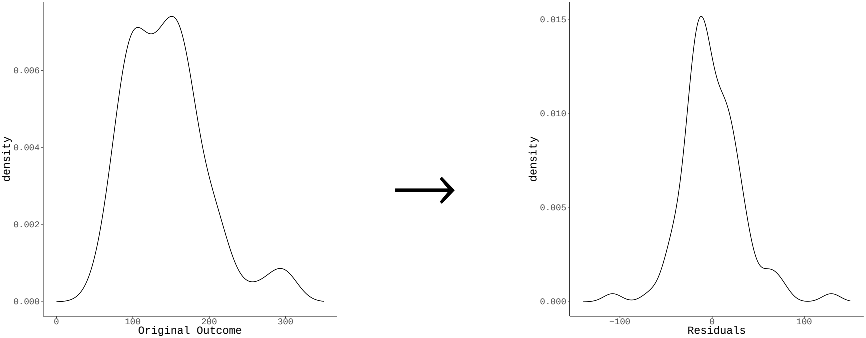
The estimated slope is: $\hat{\beta}_1 = 4.27$.

- 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.

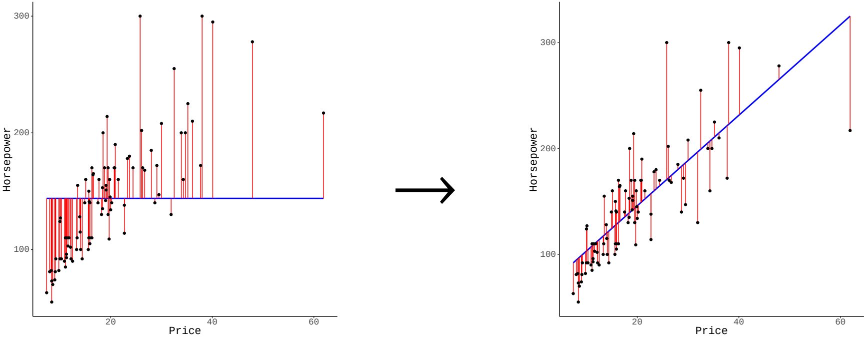
- Our model explains some proportion of the outcome's variability.
- The residual variance $\hat{\sigma}^2 = \mathrm{Var}(\hat{\varepsilon})$ will be less than $\mathrm{Var}(Y)$.



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

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

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.

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:

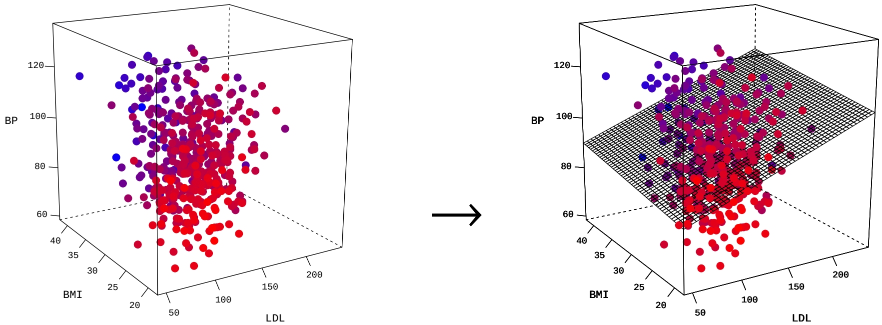
```
AIC(out1)

[1] 914.8821

BIC(out1)

[1] 922.4799
```

Multiple Linear Regression

Graphical Representations

Adding an additional predictor to a simple linear regression problem 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

```
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?*

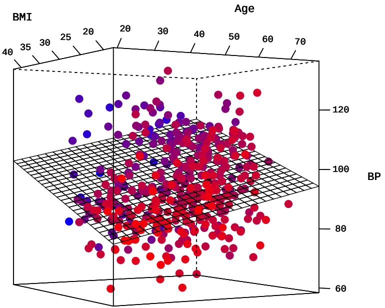
- We're partialing age out of the effect of BMI on blood pressure.

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
```

Comparing Models

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.

Comparing Models

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.

CATEGORICAL PREDICTORS

Dummy Coding

The most common way to code categorical predictors is *dummy coding*.

- A G-level factor must be converted into a set of $G - 1$ dummy codes.
- Each code is a variable on the dataset that equals 1 for observations corresponding to the code's group and equals 0, otherwise.
- The group without a code is called the *reference group*.

Example Dummy Code

Let's look at the simple example of coding biological sex:

	sex	male
1	female	0
2	male	1
3	male	1
4	female	0
5	male	1
6	female	0
7	female	0
8	male	1
9	female	0
10	female	0

Example Dummy Codes

Now, a slightly more complex example:

	drink	juice	tea
1	juice	1	0
2	coffee	0	0
3	tea	0	1
4	tea	0	1
5	tea	0	1
6	tea	0	1
7	juice	1	0
8	tea	0	1
9	coffee	0	0
10	juice	1	0

Using Dummy Codes

To use the dummy codes, we simply include the $G - 1$ codes as $G - 1$ predictor variables in our regression model.

$$Y = \beta_0 + \beta_1 X_{male} + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_{juice} + \beta_2 X_{tea} + \varepsilon$$

- The intercept corresponds to the mean of Y for the reference group.
- Each slope represents the difference between the mean of Y in the coded group and the mean of Y in the reference group.

Example

```
## Load some data:
data(Cars93, package = "MASS")

## Use a nominal predictor:
out3 <- lm(Price ~ DriveTrain, data = Cars93)
partSummary(out3, -1)

Residuals:
    Min      1Q  Median      3Q     Max
-14.050  -6.250  -1.236   3.264  32.950

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     17.63000    2.76119   6.385 7.33e-09
DriveTrainFront -0.09418    2.96008  -0.032  0.97469
DriveTrainRear  11.32000    3.51984   3.216  0.00181

Residual standard error: 8.732 on 90 degrees of freedom
Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Interpretations

- The average price of a four-wheel-drive car is $\hat{\beta}_0 = 17.63$ thousand dollars.
- The average difference in price between front-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_1 = -0.09$ thousand dollars.
- The average difference in price between rear-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_2 = 11.32$ thousand dollars.

Example

Include two sets of dummy codes:

```
out4 <- lm(Price ~ Man.trans.avail + DriveTrain, data = Cars93)
partSummary(out4, -c(1, 2))

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         21.7187     2.9222   7.432 6.25e-11
Man.trans.availYes  -5.8410     1.8223  -3.205  0.00187
DriveTrainFront     -0.2598     2.8189  -0.092  0.92677
DriveTrainRear      10.5169     3.3608   3.129  0.00237

Residual standard error: 8.314 on 89 degrees of freedom
Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Interpretations

- The average price of a four-wheel-drive car that does not have a manual transmission option is $\hat{\beta}_0 = 21.72$ thousand dollars.
- After controlling for drive type, the average difference in price between cars that have manual transmissions as an option and those that do not is $\hat{\beta}_1 = -5.84$ thousand dollars.
- After controlling for transmission options, the average difference in price between front-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_2 = -0.26$ thousand dollars.
- After controlling for transmission options, the average difference in price between rear-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_3 = 10.52$ thousand dollars.

Contrasts

All R factors have an associated *contrasts* attribute.

- The contrasts define a coding to represent the grouping information.
- Modeling functions code the factors using the rules defined by the contrasts.

```
contrasts(Cars93$Man.trans.avail)

    Yes
No    0
Yes   1
```

```
contrasts(Cars93$DriveTrain)

      Front Rear
4WD       0    0
Front     1    0
Rear      0    1
```

Significance Testing

For variables with only two levels, we can test the overall factor's significance by evaluating the significance of a single dummy code.

```
out2 <- lm(Price ~ Man.trans.avail, data = Cars93)
partSummary(out2, -c(1, 2))

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          23.841      1.623  14.691   <2e-16
Man.trans.availYes   -6.603      2.004  -3.295   0.0014

Residual standard error: 9.18 on 91 degrees of freedom
Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Significance Testing

For variables with more than two levels, we need to simultaneously evaluate the significance of each of the variable's dummy codes.

```
partSummary(out4, -c(1, 2))

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        21.7187     2.9222   7.432 6.25e-11
Man.trans.availYes  -5.8410     1.8223  -3.205  0.00187
DriveTrainFront    -0.2598     2.8189  -0.092  0.92677
DriveTrainRear     10.5169     3.3608   3.129  0.00237

Residual standard error: 8.314 on 89 degrees of freedom
Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Significance Testing

```
summary(out4)$r.squared - summary(out2)$r.squared

[1] 0.1767569

anova(out2, out4)

Analysis of Variance Table

Model 1: Price ~ Man.trans.avail
Model 2: Price ~ Man.trans.avail + DriveTrain
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     91 7668.9
2     89 6151.6  2    1517.3 10.976 5.488e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significance Testing

For models with a single nominal factor is the only predictor, we use the omnibus F-test.

```
partSummary(out3, -c(1, 2))

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     17.63000    2.76119   6.385 7.33e-09
DriveTrainFront -0.09418    2.96008  -0.032  0.97469
DriveTrainRear  11.32000    3.51984   3.216  0.00181

Residual standard error: 8.732 on 90 degrees of freedom
Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

MODEL-BASED PREDICTION

Prediction Example

To fix ideas, let's reconsider the *diabetes* data and the following model:

$$Y_{LDL} = \beta_0 + \beta_1 X_{BP} + \beta_2 X_{gluc} + \beta_3 X_{BMI} + \varepsilon$$

Training this model on the first $N = 400$ patients' data produces the following fitted model:

$$\hat{Y}_{LDL} = 22.135 + 0.089 X_{BP} + 0.498 X_{gluc} + 1.48 X_{BMI}$$

Prediction Example

To fix ideas, let's reconsider the *diabetes* data and the following model:

$$Y_{LDL} = \beta_0 + \beta_1 X_{BP} + \beta_2 X_{gluc} + \beta_3 X_{BMI} + \varepsilon$$

Training this model on the first $N = 400$ patients' data produces the following fitted model:

$$\hat{Y}_{LDL} = 22.135 + 0.089X_{BP} + 0.498X_{gluc} + 1.48X_{BMI}$$

Suppose a new patient presents with $BP = 121$, $gluc = 89$, and $BMI = 30.6$. We can predict their LDL score by:

$$\begin{aligned}\hat{Y}_{LDL} &= 22.135 + 0.089(121) + 0.498(89) + 1.48(30.6) \\ &= 122.463\end{aligned}$$

Interval Estimates for Prediction

To quantify uncertainty in our predictions, we want to use an appropriate interval estimate.

- Two flavors of interval are applicable to predictions:
    1. Confidence intervals for $\hat{Y}_m$
    2. Prediction intervals for a specific observation, Y_m

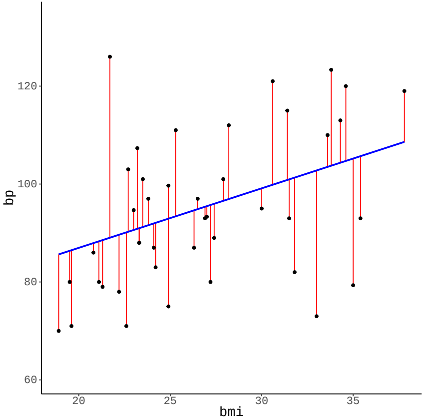
- The CI for $\hat{Y}_m$ gives a likely range (in the sense of coverage probability and “confidence”) for the *m*th value of the true conditional mean.
    - CIs only account for uncertainty in the estimated regression coefficients, $\{\hat{\beta}_0, \hat{\beta}_p\}$.

- The prediction interval for Y_m gives a likely range (in the same sense as CIs) for the *m*th outcome value.
    - Prediction intervals also account for the regression errors, ε.

Confidence vs. Prediction Intervals

Let's visualize the predictions from a simple model:

$$Y_{BP} = \hat{\beta}_0 + \hat{\beta}_1 X_{BMI} + \hat{\varepsilon}$$

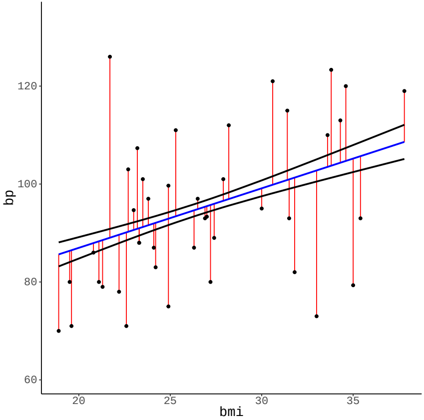


Confidence vs. Prediction Intervals

Let's visualize the predictions from a simple model:

$$Y_{BP} = \hat{\beta}_0 + \hat{\beta}_1 X_{BMI} + \hat{\varepsilon}$$

- CIs for $\hat{Y}$ ignore the errors, ε.
  - They only care about the best-fit line, $\beta_0 + \beta_1 X_{BMI}$.

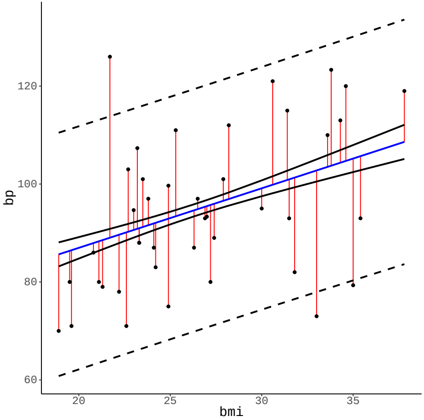


Confidence vs. Prediction Intervals

Let's visualize the predictions from a simple model:

$$Y_{BP} = \hat{\beta}_0 + \hat{\beta}_1 X_{BMI} + \hat{\varepsilon}$$

- CIs for $\hat{Y}$ ignore the errors, ε.
  - They only care about the best-fit line, $\beta_0 + \beta_1 X_{BMI}$.
- Prediction intervals are wider than CIs.
  - They account for the additional uncertainty contributed by ε.



Interval Estimates Example

Going back to our hypothetical "new" patient, we get the following 95% interval estimates:

$$95\%\ CI_{\hat{Y}} = [115.6; 129.33]$$

$$95\%\ PI = [66.56; 178.37]$$

- We can be 95% confident that the average *LDL* of patients with *Glucose* = 89, *BP* = 121, and *BMI* = 30.6 will be somewhere between 115.6 and 129.33.
- We can be 95% confident that the *LDL* of a specific patient with *Glucose* = 89, *BP* = 121, and *BMI* = 30.6 will be somewhere between 66.56 and 178.37.

MODERATION

Moderation

So far we've been discussing *additive models*.

- Additive models allow us to examine the partial effects of several predictors on some outcome.
  - The effect of one predictor does not change based on the values of other predictors.

Now, we'll discuss *moderation*.

- Moderation allows us to ask *when* one variable, X, affects another variable, Y.
  - We're considering the conditional effects of X on Y given certain levels of a third variable Z.

Equations

In additive MLR, we might have the following equation:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

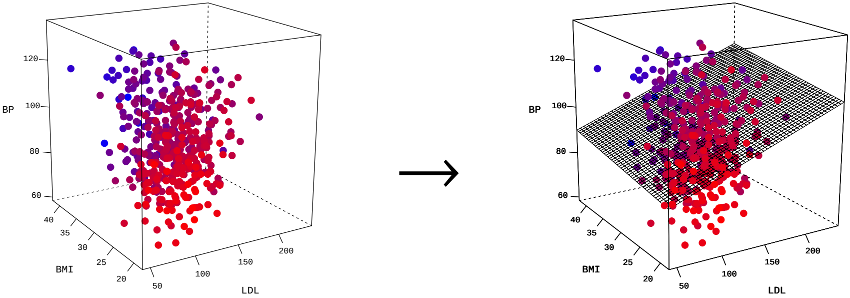
This equation assumes that X and Z are independent predictors of Y.

When X and Z are independent predictors, the following are true:

- X and Z *can* be correlated.
- β_1 and β_2 are *partial* regression coefficients.
- The effect of X on Y is the same at **all levels** of Z, and the effect of Z on Y is the same at **all levels** of X.

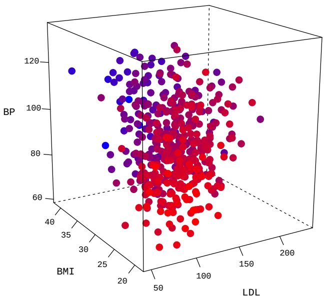
Additive Regression

The effect of X on Y is the same at **all levels** of Z.

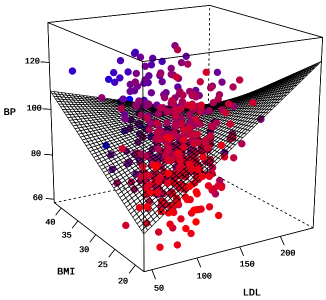


Moderated Regression

The effect of X on Y varies **as a function** of Z.



$\longrightarrow$



Equations

The following derivation is adapted from Hayes (2017).

- When testing moderation, we hypothesize that the effect of X on Y varies as a function of Z.
- We can represent this concept with the following equation:

$$Y = \beta_0 + f(Z)X + \beta_2 Z + \varepsilon \tag{1}$$

Equations

The following derivation is adapted from Hayes (2017).

- When testing moderation, we hypothesize that the effect of X on Y varies as a function of Z.
- We can represent this concept with the following equation:

$$Y = \beta_0 + f(Z)X + \beta_2 Z + \varepsilon \tag{1}$$

- If we assume that Z linearly (and deterministically) affects the relationship between X and Y, then we can take:

$$f(Z) = \beta_1 + \beta_3 Z \tag{2}$$

Equations

- Substituting Equation 2 into Equation 1 leads to:

$$Y = \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z + \varepsilon$$

Equations

- Substituting Equation 2 into Equation 1 leads to:

$$Y = \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z + \varepsilon$$

- Which, after distributing X and reordering terms, becomes:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

Testing Moderation

Now, we have an estimable regression model that quantifies the linear moderation we hypothesized.

$$\boxed{Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon}$$

- To test for significant moderation, we simply need to test the significance of the interaction term, XZ.
  - Check if $\hat{\beta}_3$ is significantly different from zero.

Interpretation

Given the following equation:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z + \hat{\beta}_3 XZ + \hat{\varepsilon}$$

- $\hat{\beta}_3$ quantifies the effect of Z on the focal effect (the $X \rightarrow Y$ effect).
  - For a unit change in Z, $\hat{\beta}_3$ is the expected change in the effect of X on Y.

- $\hat{\beta}_1$ and $\hat{\beta}_2$ are *conditional effects*.
  - Interpreted where the other predictor is zero.
  - For a unit change in X, $\hat{\beta}_1$ is the expected change in Y, when $Z = 0$.
  - For a unit change in Z, $\hat{\beta}_2$ is the expected change in Y, when $X = 0$.

Example

Still looking at the *diabetes* dataset.

- We suspect that patients' BMIs are predictive of their average blood pressure.
- We further suspect that this effect may be differentially expressed depending on the patients' LDL levels.

Example

```r
## Focal Effect:
out0 <- lm(bp ~ bmi, data = dDat)
partSummary(out0, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.9973     3.6659   16.91   <2e-16
bmi           1.2379     0.1371    9.03   <2e-16

Residual standard error: 12.72 on 440 degrees of freedom
Multiple R-squared:  0.1563,Adjusted R-squared:  0.1544
F-statistic: 81.54 on 1 and 440 DF,  p-value: < 2.2e-16
```

Example

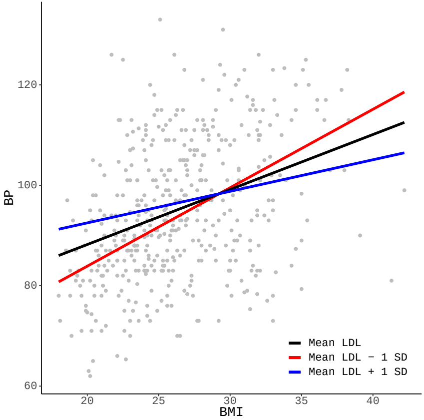
```
## Additive Model:
out1 <- lm(bp ~ bmi + ldl, data = dDat)
partSummary(out1, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.26577    3.91281  15.147  < 2e-16
bmi          1.16567    0.14156   8.235 2.08e-15
ldl          0.04016    0.02056   1.953   0.0515

Residual standard error: 12.68 on 439 degrees of freedom
Multiple R-squared:  0.1636,Adjusted R-squared:  0.1598
F-statistic: 42.94 on 2 and 439 DF,  p-value: < 2.2e-16
```

Example

```
## Moderated Model:
out2 <- lm(bp ~ bmi * ldl, data = dDat)
partSummary(out2, -c(1, 2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.480616  14.291677   1.013 0.311514
bmi          2.867825   0.541312   5.298 1.86e-07
ldl          0.448771   0.127160   3.529 0.000461
bmi:ldl     -0.015352   0.004716  -3.255 0.001221

Residual standard error: 12.54 on 438 degrees of freedom
Multiple R-squared:  0.1834,Adjusted R-squared:  0.1778
F-statistic: 32.78 on 3 and 438 DF,  p-value: < 2.2e-16
```

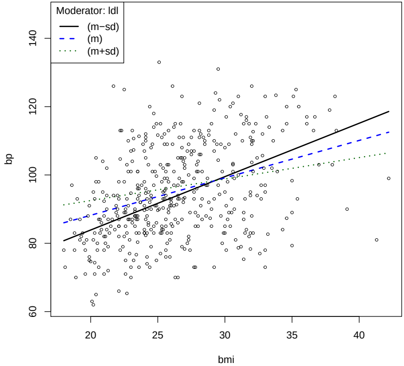
Visualizing the Interaction

We can get a better idea of the patterns of moderation by plotting the focal effect at conditional values of the moderator.



Visualizing the Interaction

```r
library(rockchalk)
plotSlopes(out2,
           plotx = "bmi",
           modx = "ldl",
           modxVals = "std.dev")
```



Categorical Moderators

Categorical moderators encode *group-specific* effects.

- E.g., if we include *sex* as a moderator, we are modeling separate focal effects for males and females.

Given a set of codes representing our moderator, we specify the interactions as before:

$$Y_{total} = \beta_0 + \beta_1 X_{inten} + \beta_2 Z_{male} + \beta_3 X_{inten} Z_{male} + \varepsilon$$

$$\begin{aligned} Y_{total} = \beta_0 &+ \beta_1 X_{inten} + \beta_2 Z_{lo} + \beta_3 Z_{mid} + \beta_4 Z_{hi} \\ &+ \beta_5 X_{inten} Z_{lo} + \beta_6 X_{inten} Z_{mid} + \beta_7 X_{inten} Z_{hi} + \varepsilon \end{aligned}$$

Example

```r
## Load data:
socSup <- readRDS(paste0(dataDir, "social_support.rds"))

## Focal effect:
out3 <- lm(bdi ~ tanSat, data = socSup)
partSummary(out3, -c(1, 2))
```

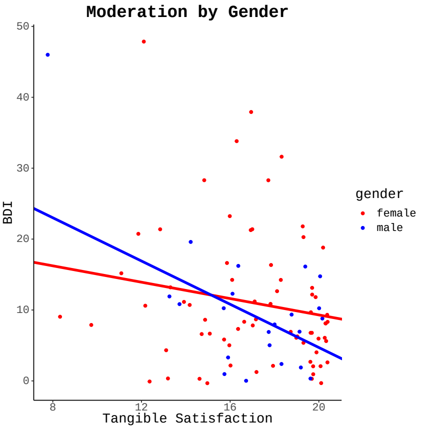
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.4089     5.3502   4.562 1.54e-05
tanSat       -0.8100     0.3124  -2.593   0.0111

Residual standard error: 9.278 on 93 degrees of freedom
Multiple R-squared:  0.06742,Adjusted R-squared:  0.05739
F-statistic: 6.723 on 1 and 93 DF,  p-value: 0.01105
```

Example

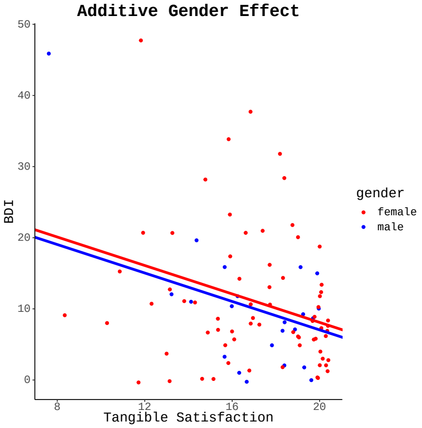
```
## Estimate the interaction:
out4 <- lm(bdi ~ tanSat * sex, data = socSup)
partSummary(out4, -c(1, 2))

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     20.8478     6.2114   3.356  0.00115
tanSat          -0.5772     0.3614  -1.597  0.11372
sexmale         14.3667    12.2054   1.177  0.24223
tanSat:sexmale  -0.9482     0.7177  -1.321  0.18978

Residual standard error: 9.267 on 91 degrees of freedom
Multiple R-squared:  0.08955,Adjusted R-squared:  0.05954
F-statistic: 2.984 on 3 and 91 DF,  p-value: 0.03537
```

Visualizing Categorical Moderation

$$\hat{Y}_{BDI} = 20.85 - 0.58X_{tsat} + 14.37Z_{male} - 0.95X_{tsat}Z_{male}$$



$$\hat{Y}_{BDI} = 28.10 - 1.00X_{tsat} - 1.05Z_{male}$$



References

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.