

# Multiple Linear Regression

Phase 3

# Key Ideas

- A Linear Regression models the relation between a dependent variable  $Y$  and an independent variable  $X$ .
- Most of the time this is not enough to explain all the variability in  $Y$ .
- We can incorporate more variables to the model → Multiple Linear Regression.
- As the number of variables increases, we may need to adjust  $R^2$ .
- For inference, we are concerned with both the entire model and the individual variables.

# Simple regression vs Multiple Linear Regression

## Simple linear regression:

Y is predicted by one variable X

## Multiple linear regression:

Y is predicted by a combination of many variables  $X_1$  ,  $X_2$ ,  $X_3$ ...

# Multiple Linear Regression

Y is predicted by a combination of many variables  $X_1, X_2, X_3 \dots$  :

$$\begin{aligned} Y_i &= \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \dots + \beta_k \cdot X_{ik} + \epsilon_i \\ &= \alpha + \sum_{j=1}^k \beta_j \cdot X_{ij} + \epsilon_i \end{aligned}$$

# Multiple Linear Regression

Y is predicted by a combination of many variables  $X_1, X_2, X_3 \dots$  :

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \dots + \beta_k \cdot X_{ik} + \epsilon_i$$
$$= \alpha + \sum_{j=1}^k \beta_j \cdot X_{ij} + \epsilon_i$$

## Notation

$\alpha$ : Intercept

$\beta_j$ : slope of the independent variable  $j$ .

$i$ : A particular observation.

$N$ : Number of observations.

$k$ : the number of independent variables.

# Practice question

An article from 1992 carried out an experiment with 30 observations to assess the impact of force, power, temperature and time on ball bond shear strength.

- What is the dependent variable?
- What is  $N$ , the number of observations?
- What is  $k$ , the number of independent variables?
- What are the independent variables?

# Assumptions

---

# Assumptions: Linear regression (Recall)

- **Linearity.** The relationship between X and Y should actually be linear!
- **Normality of residuals:** (It's actually okay if the X and Y are non-normal, as long as the residuals are normal).
- **Residuals are independent of each other.**
- **Homogeneity of variance.** Residuals generated from a normal distribution with mean 0, and with the same standard deviation for every single residual.
- **No extreme outliers:** Data points very far away from the rest can exert undue influence on the model parameters.



# Assumptions: Linear regression + no collinearity

- **Linearity.** The relationship between X and Y should actually be linear!
- **Normality of residuals:** (It's actually okay if the X and Y are non-normal, as long as the residuals are normal).
- **Residuals are independent of each other.**
- **Homogeneity of variance.** Residuals generated from a normal distribution with mean 0, and with the same standard deviation for every single residual.
- **No extreme outliers:** Data points very far away from the rest can exert undue influence on the model parameters.
- **Uncorrelated independent variables.**

# Recall: Understanding the linear regression model

- Once we have estimated the regression model, we can make **predictions** on a new X.

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} \cdot X_i$$

- Slope: For each unit increase in X, Y is expected to be **higher/lower** on average by the **slope**.

$$\hat{\beta} = r \frac{\sigma_y}{\sigma_x}$$

- Intercept: When  $X=0$  , the intercept is the expected value of Y.

$$\hat{\alpha} = \langle Y \rangle - \hat{\beta} \langle X \rangle$$

# Understanding the **multiple** linear regression model

- Once we have estimated the **multiple** regression model, we can make predictions given a new set of values for **ALL**  $X_j$ 's.

$$\hat{Y}_i = \hat{\alpha} + \sum_{j=1}^k \hat{\beta}_j \cdot X_{ij}$$

- Slopes ( $\beta_j$ ): For each unit increase in  $X_j$ ,  $Y$  is expected to be higher/lower on average by the slope, **when ALL other X variables are held constant**.
- Intercept ( $\alpha$ ): the expected value of  $Y$ , **when ALL the  $X_j=0$** .

## Practice question

A trucking company considered a multiple regression model to relate the total daily travel hours of their drivers ( $Y$ ) to the distance travel in miles ( $X_1$ ) and the number of deliveries made ( $X_2$ ). Suppose that the estimated regression model is:

$$Y = -0.8 + 0.06 \cdot X_1 + 0.9 \cdot X_2 + \epsilon$$

- A) What is the expected travel time when 50 miles are traveled and three delivered are made?
- B) How would you interpret the coefficients 0.06 and 0.9?

# Estimation: Ordinary Least Squares

Like the usual linear regression model, we want to estimate the intercept  $\alpha$  and slopes  $\beta_j$  such that the residuals are as small as possible

$$\sum_i \epsilon_i^2 = \sum_i (Y - \hat{Y}_i)^2$$

# Estimation: Ordinary Least Squares

Like the usual linear regression model, we want to estimate the intercept  $\alpha$  and slopes  $\beta_j$  such that the residuals are as small as possible

$$\sum_i \epsilon_i^2 = \sum_i (Y - \hat{Y}_i)^2$$

$$\hat{\beta} \equiv (\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k)$$



$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

$$\begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 1 & X_{N1} & \dots & X_{Nk} \end{bmatrix}$$


## Performance: Coefficient of determination $R^2$

- We saw previously that the performance of the fit of a regression model can be evaluated using the coefficient of determination,  $R^2$ .
- Formally, it is defined as the part of variability explained by our predictors out of the total variability.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

## Performance: Coefficient of determination $R^2$

- We saw previously that the performance of the fit of a regression model can be evaluated using the coefficient of determination,  $R^2$ .
- Formally, it is defined as the part of **variability explained by our predictors** out of the total variability.


$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$\sum_i (Y_i - \hat{Y}_i)^2$$



## Performance: Coefficient of determination $R^2$

- We saw previously that the performance of the fit of a regression model can be evaluated using the coefficient of determination,  $R^2$ .
- Formally, it is defined as the part of variability explained by our predictors out of the **total variability**.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

  $\sum_i (Y_i - \langle Y \rangle)^2$

# Performance: Coefficient of determination $R^2$

- We saw previously that the performance of the fit of a regression model can be evaluated using the coefficient of determination,  $R^2$ .
- Formally, it is defined as the part of **variability explained by our predictors** out of the **total variability**.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Diagram illustrating the components of the coefficient of determination formula:

- The numerator  $SS_{res}$  (Residual Sum of Squares) is linked by a red arrow to the formula  $\sum_i (Y_i - \hat{Y}_i)^2$ .
- The denominator  $SS_{tot}$  (Total Sum of Squares) is linked by a red arrow to the formula  $\sum_i (Y_i - \langle Y \rangle)^2$ .

# Performance: Coefficient of determination $R^2$

- We saw previously that the performance of the fit of a regression model can be evaluated using the coefficient of determination,  $R^2$ .
- Formally, it is defined as the part of variability explained by our predictors out of the total variability.
- Properties:
  - If our predictors explain all the variability of  $Y$ ,  $SS_{\text{res}} = 0 \rightarrow R^2=1$
  - If our predictors explained none variability of  $Y$ ,  $SS_{\text{res}} = SS_{\text{tot}} \rightarrow R^2=0$

## Performance: Coefficient of determination $R^2$

- We saw previously that the performance of the fit of a regression model can be evaluated using the coefficient of determination,  $R^2$ .
- Formally, it is defined as the part of variability explained by our predictors out of the total variability.

$$R^2 = 1 - \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \langle Y \rangle)^2}$$

# Adjusted coefficient of determination $R^2$

- There is one issue with the previous formula: as we add more predictors into the model,  $R^2$  will ALWAYS increase (or at least not decrease)
- As a result, for multiple linear regression, it's common to adjust  $R^2$  to take into account the number of predictors.

$$\text{adj. } R^2 = 1 - \left( \frac{SS_{res}}{SS_{tot}} \times \frac{N - 1}{N - k - 1} \right)$$

# Adjusted coefficient of determination $R^2$

- There is one issue with the previous formula: as we add more predictors into the model,  $R^2$  will ALWAYS increase (or at least not decrease)
- As a result, for multiple linear regression, it's common to adjust  $R^2$  to take into account the number of predictors.
- Property: If we add a variable that does not really provide any new information - or is completely unrelated- the adjusted  $R^2$  does not increase.

# Adjusted coefficient of determination $R^2$

- There is one issue with the previous formula: as we add more predictors into the model,  $R^2$  will ALWAYS increase (or at least not decrease)
- As a result, for multiple linear regression, it's common to adjust  $R^2$  to take into account the number of predictors.
- Property: If we add a variable that does not really provide any new information - or is completely unrelated- the adjusted  $R^2$  does not increase.
- Limitation: Not clear interpretation like  $R^2$  (cannot interpret as a variability explained portion).

# Inference

- We've talked about estimating the regression model and its performance. But how do all these relate to our assumed model (i.e. null hypothesis)?
- Here we have two different (but related) kinds of null hypothesis testing:
  - 1- Does our regression model perform significantly better 0?
  - 2 - Which independent variables significantly contribute to this?
- Example: Is anxiety significantly explained by the amount of time spent exercising and of sleep a person? And if so, which of these is significant?
- Analogy: ANOVA and t-tests for post-hoc analysis.



# Inference

- We've talked about estimating the regression model and its performance. But how do all these relate to our assumed model (i.e. null hypothesis)?
- Here we have two different (but related) kinds of null hypothesis testing:
  - 1- Does our regression model perform significantly better 0?
  - 2 - Which independent variables significantly contribute to this?
- Example: Is anxiety significantly explained by the amount of time spent exercising and of sleep a person? And if so, which of these is significant?
- Analogy: ANOVA and t-tests for post-hoc analysis. **Indeed, all come down to regression in the end... → Next week!**

# Inference: full model

➤ Null Hypothesis  $\mathbf{H}_0$ :  $\hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_k = 0$

➤ Test statistic:  $F = \frac{(SS_{tot} - SS_{res})/k}{SS_{res}/N - k - 1}$

➤ Alternative Hypothesis  $\mathbf{H}_A$ :

At least one  $\hat{\beta}_j \neq 0 \quad (j = 1 \dots k)$

## Recall: one way ANOVA test (Week 7)

- Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_j$
- Test statistic:  $F = \frac{V_b / df_b}{V_w / df_w} \sim \text{F-distribution } (df_b, df_w)$
- Alternative Hypothesis  $H_A$   
at least two of the  $\mu_j$ 's are different

# Inference: full model

➤ Null Hypothesis  $\mathbf{H}_0$ :  $\hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_k = 0$

➤ Test statistic:  $F = \frac{(SS_{tot} - SS_{res})/k}{SS_{res}/N - k - 1} \sim \text{F-distribution}(k, N - k - 1)$

➤ Alternative Hypothesis  $\mathbf{H}_A$ :


At least one  $\hat{\beta}_j \neq 0 \quad (j = 1 \dots k)$

# Inference: full model

➤ Null Hypothesis  $\mathbf{H}_0$ :  $\hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_k = 0$

➤ Test statistic:  $F = \frac{(SS_{tot} - SS_{res})/k}{SS_{res}/N - k - 1} \sim \text{F-distribution}(k, N - k - 1)$

Degrees of freedom of the **model**.



➤ Alternative Hypothesis  $\mathbf{H}_A$ :

At least one  $\hat{\beta}_j \neq 0 \quad (j = 1 \dots k)$

# Inference: full model

➤ Null Hypothesis  $\mathbf{H}_0$ :  $\hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_k = 0$

➤ Test statistic:  $F = \frac{(SS_{tot} - SS_{res})/k}{SS_{res}/N - k - 1} \sim \text{F-distribution}(k, N - k - 1)$

Degrees of freedom of the  
**residuals**  
(question: Why also this -1?)

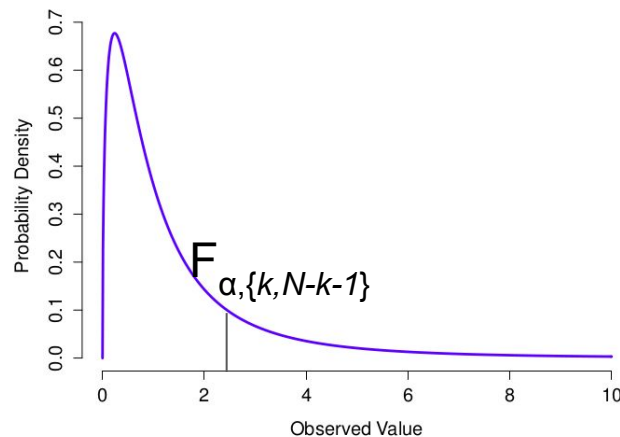
➤ Alternative Hypothesis  $\mathbf{H}_A$ :

At least one  $\hat{\beta}_j \neq 0$  ( $j = 1 \dots k$ )

# Inference: full model

➤ Null Hypothesis  $\mathbf{H}_0$ :  $\hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_k = 0$

➤ Test statistic:  $F = \frac{(SS_{tot} - SS_{res})/k}{SS_{res}/N - k - 1}$



➤ Alternative Hypothesis  $\mathbf{H}_A$ :

Rejection region for  $\alpha$

At least one  $\hat{\beta}_j \neq 0 \quad (j = 1 \dots k) \quad F \geq F_{\alpha, \{k, N-k-1\}}$

# Inference: full model

➤ Null Hypothesis  $\mathbf{H}_0$ :  $\hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_k = 0$

➤ Test statistic:  $F = \frac{(SS_{tot} - SS_{res})/k}{SS_{res}/N - k - 1}$

`lm(formula, data)`

➤ Alternative Hypothesis  $\mathbf{H}_A$ : Rejection region for  $\alpha$  (in P-VALUES)

At least one  $\hat{\beta}_j \neq 0$  ( $j = 1 \dots k$ )  $P(f > F \mid H_0)$



# Practice question

An article from 1992 carried out an experiment with 30 observations to assess the impact of force, power, temperature and time on ball bond shear strength. A multiple linear regression model was estimated, with the following results:

$$SS_{\text{tot}} = 2325.2587$$

$$SS_{\text{res}} = 665.1187$$

- A) How much variability of ball bond shear strength do the four variables explain?
- B) Is this variability statistically greater than 0 at a significance level of 0.05?  
(Hint:  $F_{4, 25} = 6.49$ )

# Beyond the full model: inference for each predictor

➤ Null Hypothesis  $H_0$ :  $\beta_j = 0$

➤ Test statistic:  $t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$

➤ Alternative Hypothesis  $H_A$ :

$\beta_j > 0$  (one-sided right tail)

$\beta_j < 0$  (one-sided left tail)

$\beta_j \neq 0$  (two-sided)

# Beyond the full model: inference for each predictor

➤ Null Hypothesis  $H_0$ :  $\beta_j = 0$

➤ Test statistic:  $t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim \text{Student's } t (N-k-1)$

➤ Alternative Hypothesis  $H_A$ :

$\beta_j > 0$  (one-sided right tail)

$\beta_j < 0$  (one-sided left tail)

$\beta_j \neq 0$  (two-sided)

# Beyond the full model: inference for each predictor

➤ Null Hypothesis  $H_0: \beta_j = 0$

➤ Test statistic:  $t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim \text{Student's } t (N-k-1)$

➤ Alternative Hypothesis  $H_A$ : Rejection region for  $\alpha$

$\beta_j > 0$  (one-sided right tail)

$$t \geq t_{\alpha, N-k-1}$$

$\beta_j < 0$  (one-sided left tail)

$$t \leq t_{\alpha, N-k-1}$$

$\beta_j \neq 0$  (two-sided)

$$|t| \geq |t_{\alpha/2, N-k-1}|$$

# Beyond the full model: inference for each predictor

➤ Null Hypothesis  $H_0: \beta_j = 0$

➤ Test statistic:  $t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim \text{Student's } t (N-k-1)$

➤ Alternative Hypothesis  $H_A$ : Rejection region for  $\alpha$  (in P-VALUES)

$\beta_j > 0$  (one-sided right tail)

$$P(T \geq t \mid H_0) \leq \alpha$$

$\beta_j < 0$  (one-sided left tail)

$$P(T \leq t \mid H_0) \leq \alpha$$

$\beta_j \neq 0$  (two-sided)

$$P(T \geq |t| \mid H_0) \leq \alpha$$

# Beyond the full model: inference for each predictor

➤ Null Hypothesis  $H_0: \beta_j = 0$

➤ Test statistic:  $t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim \text{Student's } t (N-k-1)$

`lm(formula, data)`

R gives the p-values for each predictor in addition to the full model!!!

➤ Alternative Hypothesis  $H_A$ : Rejection region for  $\alpha$  (in P-VALUES)

$\beta_j > 0$  (one-sided right tail)

$$P(T \geq t \mid H_0) \leq \alpha$$

$\beta_j < 0$  (one-sided left tail)

$$P(T \leq t \mid H_0) \leq \alpha$$

$\beta_j \neq 0$  (two-sided)

$$P(T \geq |t| \mid H_0) \leq \alpha$$

## Practice question (Continuation)

An article from 1992 carried out an experiment with 30 observations to assess the impact of force, power, temperature and time on ball bond shear strength. A multiple linear regression model was estimated, with the following results for each regression coefficient:

C) Which independent variables have a significant association ( $\neq 0$ ) with ball bond shear strength at a usual 0.05 significance level? (Hint:  $t_{0.025,25} = 2.06$ )

Coefficient	Estimate	SE
intercept	-37.47667	13.09964
force	0.2116667	0.210574
power	0.4983333	0.070191
temp	0.1296667	0.042115
time	0.2583333	0.210574

# Confidence intervals for the coefficients

- Like in any other parameter estimation, estimated regression coefficients will exhibit some **variability** with respect to the true value.
- We can easily build the  $100*(1-\alpha)\%$  confidence intervals around these as follows:

$$\begin{aligned}\hat{\alpha} &\pm t_{\alpha/2, N-k-1} \times SE(\hat{\alpha}) \\ \hat{\beta}_j &\pm t_{\alpha/2, N-k-1} \times SE(\hat{\beta}_j)\end{aligned}$$



# Confidence intervals for the coefficients

- Like in any other parameter estimation, estimated regression coefficients will exhibit some **variability** with respect to the true value.
- We can easily build the  $100*(1-\alpha)\%$  confidence intervals around these as follows:

R:  
`confint(lm.model, level = (1- $\alpha$ ))`

$$\hat{\alpha} \pm t_{\alpha/2, N-k-1} \times SE(\hat{\alpha})$$
$$\hat{\beta}_j \pm t_{\alpha/2, N-k-1} \times SE(\hat{\beta}_j)$$

# Regression and categorical variables

- All the previous example used **continuous independent variables**.
- In this case: any **change** in the independent variable triggers a change in the dependent variable **proportional** with the **slope**  $\square$ .
- How about categorical variables? How are these changes defined?
- In these cases, categorical variables are usually encoded using **dummy variables**.
- A **dummy variable** concentrates on **one** particular **level** of a categorical variable, with a value 1 if that observation has that particular level, and 0 otherwise.

## Dummy variable: example

The article “Estimating Urban Travel Times: A Comparative Study” (Trans. Res., 1980: 173–175) described a study relating the dependent variable  $Y$ , travel time between locations in a certain city, and the independent variable  $X_2$ , the distance between locations. There was another variable  $X_1$ : types of vehicle (cars or truck).

# Dummy variable: example

The article “Estimating Urban Travel Times: A Comparative Study” (Trans. Res., 1980: 173–175) described a study relating the dependent variable  $Y$ , travel time between locations in a certain city, and the independent variable  $X_2$ , the distance between locations. There was another variable  $X_1$ : types of vehicle (cars or truck).

➤ **Dummy variable encoding** of  $X_1$ : 1 if a truck; 0 if a car.

# Dummy variable: example

The article “Estimating Urban Travel Times: A Comparative Study” (Trans. Res., 1980: 173–175) described a study relating the dependent variable  $Y$ , travel time between locations in a certain city, and the independent variable  $X_2$ , the distance between locations. There was another variable  $X_1$ : types of vehicle (cars or truck).

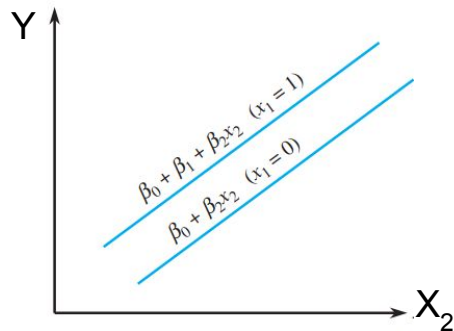
- **Dummy variable encoding** of  $X_1$ : 1 if a truck; 0 if a car.
- The regression model is still **the same**:  $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

# Dummy variable: example

The article “Estimating Urban Travel Times: A Comparative Study” (Trans. Res., 1980: 173–175) described a study relating the dependent variable  $Y$ , travel time between locations in a certain city, and the independent variable  $X_2$ , the distance between locations. There was another variable  $X_1$ : types of vehicle (cars or truck).

- **Dummy variable encoding** of  $X_1$ : 1 if a truck; 0 if a car.
- The regression model is still **the same**:  $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- **One regression line for each level:**

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_2 X_2 \quad (\text{Car})$$
$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 X_2 \quad (\text{Truck})$$



# Dummy variable: example

The article “Estimating Urban Travel Times: A Comparative Study” (Trans. Res., 1980: 173–175) described a study relating the dependent variable  $Y$ , travel time between locations in a certain city, and the independent variable  $X_2$ , the distance between locations. There was another variable  $X_1$ : types of vehicle (cars or truck).

- **Dummy variable encoding** of  $X_1$ : 1 if a truck; 0 if a car.
- The regression model is still **the same**:  $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- **One regression line for each level:**

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_2 X_2 \quad (\text{Car})$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 X_2 \quad (\text{Truck})$$

## Interpretation

$\hat{\beta}_1$ : difference in average travel time between trucks and cars.

$\hat{\beta}_1 > 0 \rightarrow$  trucks have longer travel times than cars.

# Recap

- Most of the time a simple Linear Regression this is not enough to explain all the variability in  $Y$ .
- We can incorporate more variables to the model → Multiple Linear Regression.
- As the number of variables increase, we may need to adjust  $R^2$ .
- For inference, we are concerned with both the entire model and the individual variables.