Lecture 4:
Multiple Linear Regression
STAT 452, Spring 2021

Multiple Linear Regression (MLR)

Suppose y is a response variable, and $x_1, \cdots, x_p$ are p explanatory variables. Then the multiple linear regression model can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$ is the random error term.

Multiple Linear Regression (MLR)

Suppose we have a collection $i = 1, \cdots, n$ observations. Then the multiple linear regression model for case i is written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \cdots + \beta_p x_{ip} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently.

Given estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ of the unknown regression parameters $\beta_0, \beta_1, \cdots, \beta_p$:

- The i^{th} fitted (or predicted) value:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$$

- The i^{th} residual:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}$$

Least Squares Estimation

The parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ are found by minimizing the sum of squared residuals:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

- Using some calculus, a closed form solution for the parameter estimates can be derived and expressed using matrix notation.
- In practice, for a specific data set, we can use the `lm()` function in R to compute the least squares estimates of the parameters.

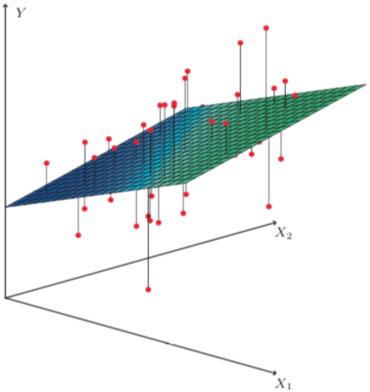


FIGURE 3.4. *In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.*

From Chapter 3, p. 73, of *An Introduction to Statistical Learning*.

Hypothesis Test for a Single Predictor

Test whether parameter β_j is zero.

$H_0 : \beta_j = 0$
$H_A : \beta_j \neq 0$

Test statistic:

$$t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}; \quad \text{df} = n - p - 1$$

- $se(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$
- n is the number of observations
- p is the number of predictor variables
- degrees of freedom (df) $=$
  sample size - number of parameters estimated $= n - p - 1$
  (since, when including the intercept, there are $p + 1$ parameters)

Confidence Interval for a Single Predictor

A $1-\alpha$ confidence interval for β_j:

$$\hat{\beta}_j \pm t_{\alpha/2;n-p-1} se(\hat{\beta}_j)$$

The R function `confint()` can be used to calculate confidence intervals for the parameters.

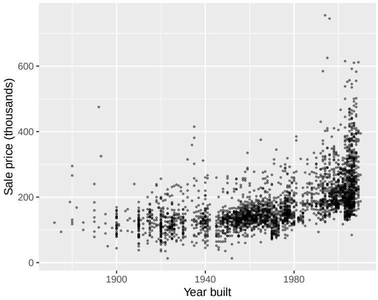
Example: Ames Housing Data

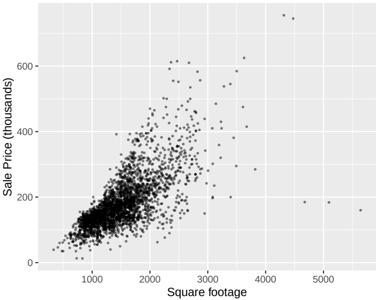
- We again use a data set on residential properties in Ames, Iowa, which can be accessed though the R package `AmesHousing`

- We will consider the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

  - y is `Sale_Price`, the sale price in dollars
  - x_1 is `Gr_Liv_Area`, the above ground living area in square feet
  - x_2 is `Year_Built`, the year the property was built


Sale price (thousands)
600
400
200
0
1900
1940
1980
Year built


Sale Price (thousands)
600
400
200
0
1000
2000
3000
4000
5000
Square footage

```
> library(AmesHousing)
> ames <- make_ames()

# set global R options
> options(scipen = 10)

> lm2 <- lm(Sale_Price ~ Gr_Liv_Area + Year_Built, data = ames)
> summary(lm2)

Coefficients:
                Estimate   Std. Error t value Pr(>|t|)
(Intercept) -2106459.470    57338.740  -36.74   <2e-16 ***
Gr_Liv_Area       95.969        1.758   54.60   <2e-16 ***
Year_Built      1087.237       29.377   37.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46660 on 2927 degrees of freedom
Multiple R-squared:  0.6591,Adjusted R-squared:  0.6588
F-statistic:  2829 on 2 and 2927 DF,  p-value: < 2.2e-16


> confint(lm2, level = 0.95)
                     2.5 %          97.5 %
(Intercept) -2218887.82640 -1994031.11360
Gr_Liv_Area       92.52311       99.41588
Year_Built      1029.63523     1144.83823
```

The equation for the estimated regression model:

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ &= -2106459.47 + 95.97 x_1 + 1087.24 x_2\end{aligned}$$

where $\hat{y}$ is the prediction for `Sale_Price`, x_1 is `Gr_Liv_Area`, and x_2 is `Year_Built`.

R code to predict `Sale_Price` when `Gr_Liv_Area` $= 1500$ and `Year_Built` $= 1990$:

```
> new_x <- data.frame(Gr_Liv_Area = 1500, Year_Built = 1990)
> predict(lm2, newdata = new_x)
       1
201095.9
```

Example: Your Turn

(a) Interpret $\hat{\beta}_2$, the estimated coefficient for `Year_Built`

(b) Interpret the coefficient of determination (R^2).

(c) Use R to predict the average sales price for a property built in the year 1970, and with 1100 square feet of above ground living area.