Lecture 3: Simple Linear Regression

Module 1: part 2

Spring 2024

Logistics

- Lecture today will introduce linear regression with 1 covariate
- Labs Mon/Tues will cover fitting linear models in `R`
- Module 1 assessment will be posted before Monday Feb 3, due date is Tues Feb 9, 11:59pm

Recap

- Population $\to$ Data $\to$ Statistic
- Mean, median, mode can be seen as minimizing certain criteria
- Correlation measures linear association between two variables

Linear regression

Parameters which govern a line

The equation for a line can be put into the following form

$$Y = b_0 + b_1 X \tag{1}$$

Parameters which govern a line

The equation for a line can be put into the following form

$$Y = b_0 + b_1 X \tag{1}$$

- X and Y are variables
- b_0 is the **Y-intercept**. It is the value of the Y coordinate when X = 0
- b_1 is the **slope**. It describes how Y changes as X changes.

Wine vs Ratings

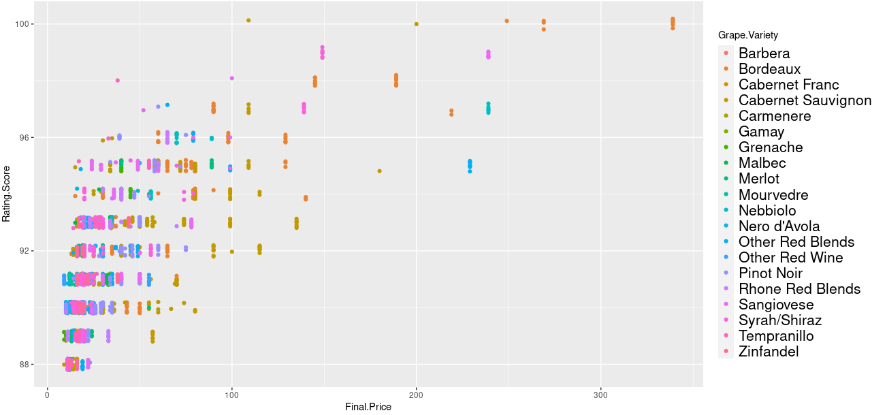


Figure: Data from Wine.com circa 2015

Alternative way

Summarize a set of numbers $\{2, 5, 8, 10\}$

- Let $\hat{b}_0$ be a "candidate"
- The residual for x_i is $x_i - \hat{b}_0$
- Measure how well the candidate summarizes the set by the *residual sum of squares*

$$RSS(\hat{b}_0) = \sum_i |x_i - \hat{b}_0|^2 = \sum_i |e_i|^2$$

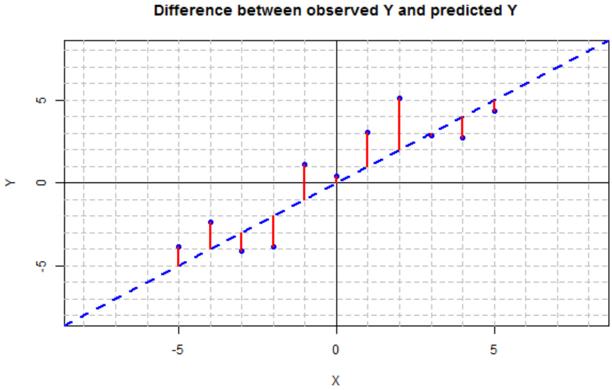
- Suppose $\hat{b}_0 = 6$

x_i	$x_i - \hat{b}_0$	$(x_i - \hat{b}_0)^2$
2	-4	16
5	-1	1
8	2	4
10	4	16

Errors in Y

Suppose we observe $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. To select a "best line" we consider the difference between the predicted point and observed value of y_i.

Predicted value: $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$
Residual: $e_i = y_i - \hat{y}_i$



Selecting Regression Coefficient

How can we select a slope and intercept to minimize the sum of squared errors?

$$RSS(\hat{b}_0, \hat{b}_1) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2 \quad (2)$$

Selecting Regression Coefficient

How can we select a slope and intercept to minimize the sum of squared errors?

$$RSS(\hat{b}_0, \hat{b}_1) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2 \tag{2}$$

Next few slides have math, which you can look through more carefully if you want, but is otherwise not necessary

Selecting Regression Coefficient

How can we select a slope and intercept to minimize the sum of squared errors?

$$RSS(\hat{b}_0, \hat{b}_1) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2 \tag{2}$$

Next few slides have math, which you can look through more carefully if you want, but is otherwise not necessary

Take a derivative and set equal to 0!

$$\frac{\partial RSS}{\partial \hat{b}_1} = -2 \sum_i x_i (y_i - (\hat{b}_0 + \hat{b}_1 x_i)) = 0 \tag{3}$$

$$\frac{\partial RSS}{\partial \hat{b}_0} = -2 \sum_i (y_i - (\hat{b}_0 + \hat{b}_1 x_i)) = 0 \tag{4}$$

Selecting Regression Coefficient: $\hat{b}_0$

$$
\begin{aligned}
0 = \frac{\partial RSS}{\partial \hat{b}_0} &= -2\sum_i (y_i - (\hat{b}_0 + \hat{b}_1 x_i)) \\
&= -2\sum_i y_i + 2n\hat{b}_0 + \hat{b}_1 \sum_i x_i
\end{aligned} \tag{5}
$$

$$
\Rightarrow \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} \tag{6}
$$

Selecting Regression Coefficient: $\hat{b}_1$

$$\begin{aligned}\frac{\partial RSS}{\partial \hat{b}_1} &= -2\sum_i x_i(y_i - (\hat{b}_0 + \hat{b}_1 x_i)) = 0 \\ &= \sum_i x_i(y_i - (\hat{b}_0 + \hat{b}_1 x_i))\end{aligned} \quad (7)$$

$$\hat{b}_1 = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{(n-1)}\sum_i (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x} \quad (8)$$

Ordinary least squares regression

This procedure is called Ordinary least squares (OLS) or simple linear regression

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i \tag{9}$$

- The best fit line passes through the centroid $(\bar{x}, \bar{y})$
- $y_i - \hat{y}_i$ is called the **residual**
- The sum of the residuals for the best fit line is 0
- We say Y is “regressed onto” X
- The estimated parameters are not symmetric. If we swap what is “x” and what is “y”, the line will change.

Ordinary least squares regression

This procedure is called Ordinary least squares (OLS) or simple linear regression

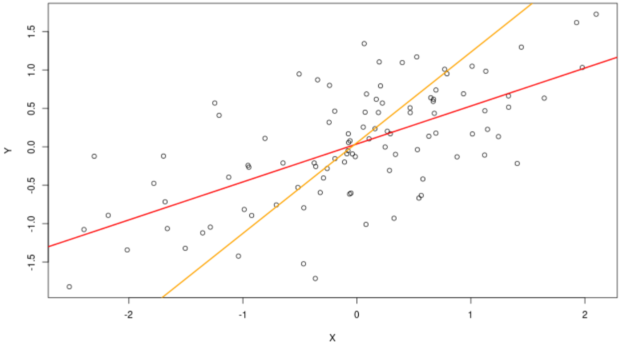


Figure: Red is Y regressed onto X; Orange is X regressed onto Y

Outliers

Outliers and Influential Points

You will see in the lab next week, that an outlier can drastically effect the results of a regression.

Outliers and Influential Points

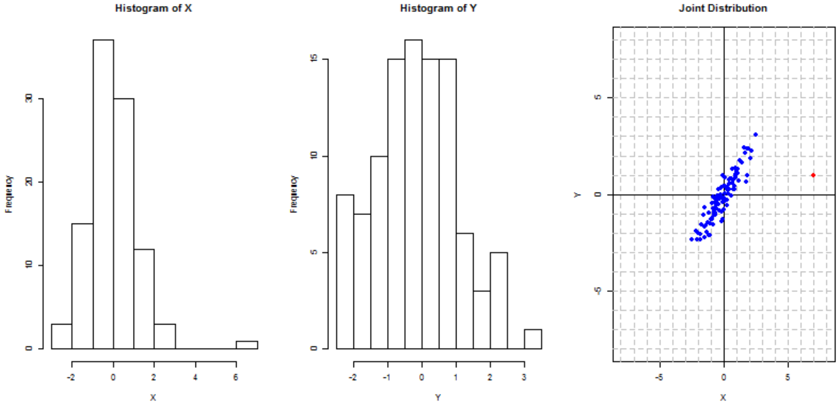
You will see in the lab next week, that an outlier can drastically effect the results of a regression.

Outliers are "unusual" observations. But what does it mean to be "unusual"?

- Unusual X value (marginal)
- Unusual Y value (marginal)
- Unusual X and Y value together (joint)
- Might be consistent with the trend, might be inconsistent with the trend

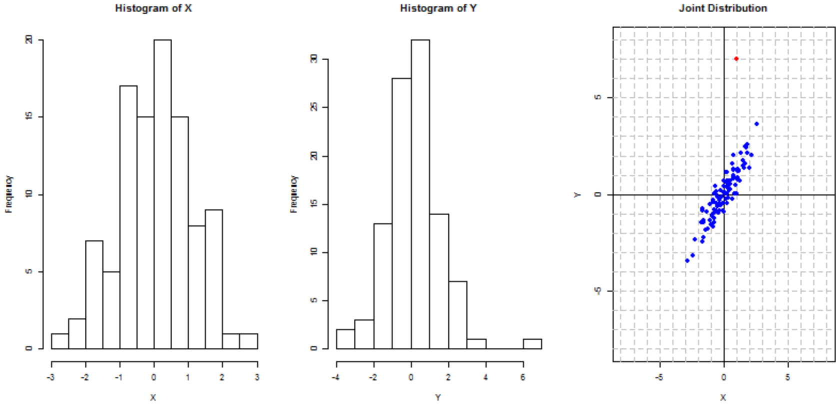
Outliers and Influential Points

Unusual X Value



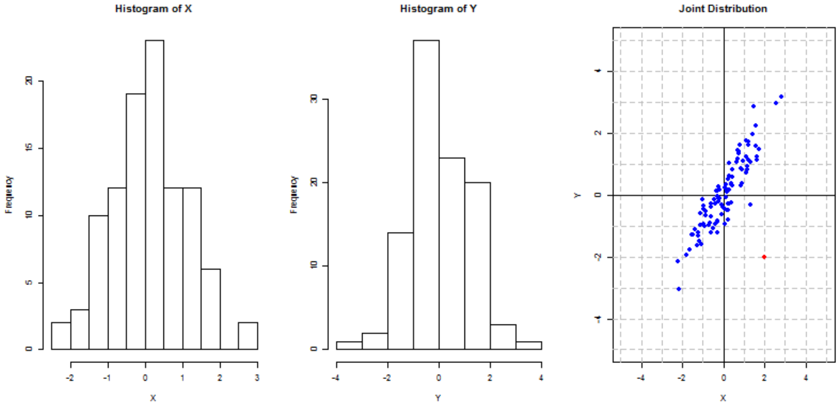
Outliers and Influential Points

Unusual Y Value



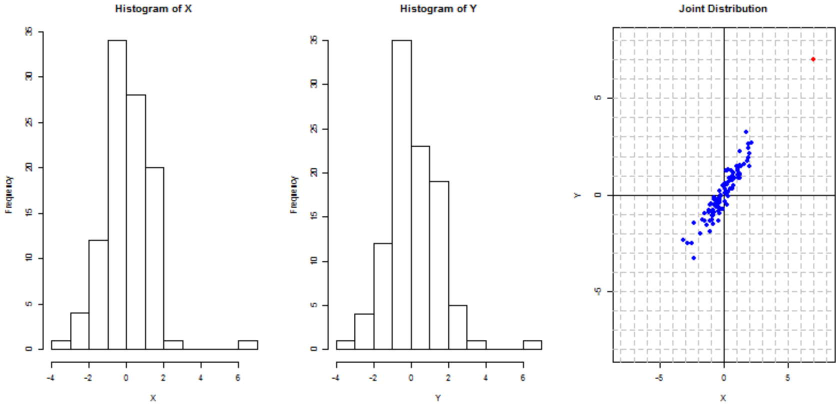
Outliers and Influential Points

Unusual X and Y Value, inconsistent with the trend



Outliers and Influential Points

Unusual X and Y Value, but consistent with the trend



Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b}_1 = cov(X, Y)/var(X) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad (10)$$

Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b}_1 = cov(X, Y)/var(X) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{10}$$

Does $\hat{b}_1$ change if I add a point at

- $(\bar{x}, \bar{y})$.

Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b}_1 = cov(X, Y)/var(X) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{10}$$

Does $\hat{b}_1$ change if I add a point at

- $(\bar{x}, \bar{y})$. No!
- $(\bar{x}, y)$.

Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b}_1 = cov(X, Y)/var(X) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{10}$$

Does $\hat{b}_1$ change if I add a point at

- $(\bar{x}, \bar{y})$. No!
- $(\bar{x}, y)$. No!
- $(x, \bar{y})$.

Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b}_1 = cov(X, Y)/var(X) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{10}$$

Does $\hat{b}_1$ change if I add a point at

- $(\bar{x}, \bar{y})$. No!
- $(\bar{x}, y)$. No!
- $(x, \bar{y})$. Yes!

Outliers in the X direction affect the slope much more than outliers in the Y direction

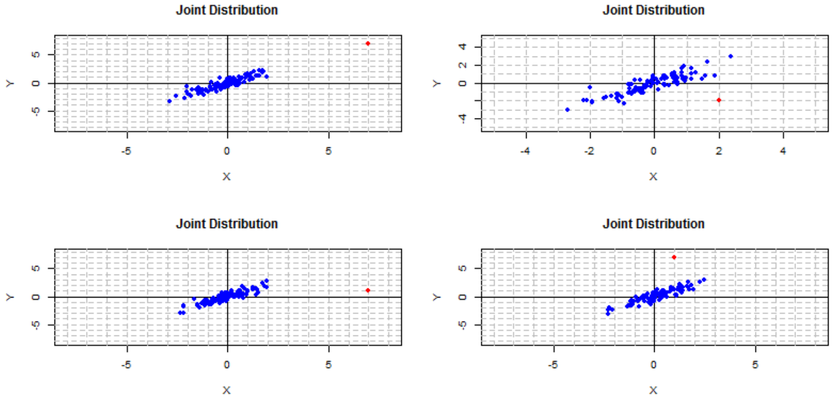
Outliers and Influential Points

Outliers in the X direction can affect the slope much more than outliers in the Y direction

- Leverage- Points where the x_i is far from $\bar{x}$ have high leverage
- Influence- Points whose inclusion/exclusion drastically change the regression slope. High leverage can increase influence. Depends on both X and Y values

Outliers and Influential Points

Are the previous outliers we showed high leverage? high influence?



Other estimators

- We motivated "Least Squares" regression as selecting the line (or the parameters of the line) which minimizes the RSS of the data
- But recall we could have defined other estimators (L_1)

$$(\hat{b}_0, \hat{b}_1) = \arg\min_{b_0, b_1} \sum_i |y_i - (b_0 - b_1 x_i)| \tag{11}$$

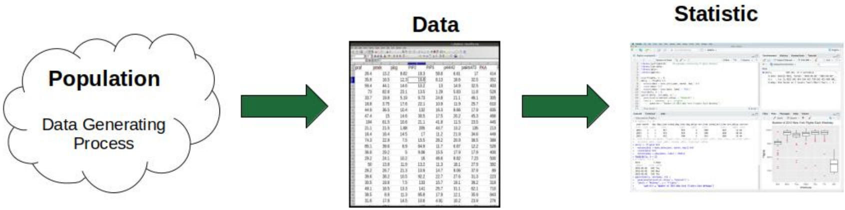
- The analogue of a "median line"
- Less affected by outliers
- Least absolute deviation estimator is a special case of what is called quantile regressions
- Will see example in lab

Outliers and Influential Points

So what should we do with outliers?

- As with most thing in statistics... it depends
- What do we know about the outlier? What trend are we trying to capture?

Sample data vs population distribution



Wrap-up

- Introduce linear regression as procedure which minimizes the squared residuals
- Gave intuition for how outliers might effect resulting regression estimates