Stat 177

Introduction to regression models

Richard P. Waterman

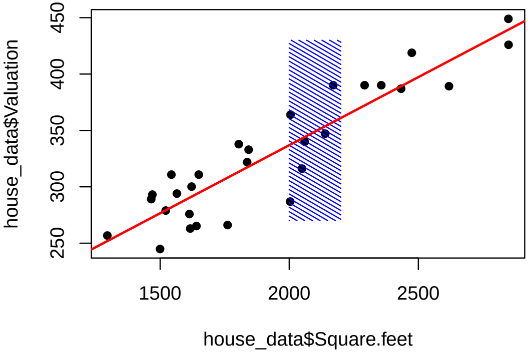
Wharton

Summer 2020

Table of contents I

Motivating example

- Using the real estate data set (`real_housing.csv`), what are reasonable approaches to estimating the value of a house that has 2100 square feet of space?
- What about a house with 5000 square feet?



Two approaches

1. Be local: use only information in the neighborhood of the prediction.
2. Be global: assert a universal truth (like linearity) and exploit that assertion. As lines have constant slope, if we believe in linearity, then we can use all the data to estimate that common slope.

- Local approach:
  - Benefit: makes fewer assumptions.
  - Downsides: potentially leaves some information *on the table*. Doesn't work so well in high dimensions as the neighborhoods are sparsely populated.
- Global approach:
  - Benefit: uses all of the data for more precision.
  - Downside: if the global assumption is wrong, then it's a fool's paradise.

Notation for lines in the notes

We will use the notation:

$$y = b_0 + b_1 x,$$

for the least squares line

- y is the *response variable*
- x is the *predictor variable*
- b_0 is the intercept
- b_1 is the slope

The defining property of a line: the slope is constant.

Reasons to fit a line through data

Once we have an equation we can summarize and exploit fit:

- Graphically summarize.
- Prediction:
  - Interpolate.
  - Forecast/extrapolate (with caution).
- Answer questions: in an adjacent area, the incremental value per square foot is \$100. Does this neighborhood appear different?
- How much of the variation in valuation does this model account for?
- How precise are the predictions from this model?
- If a 2500 square foot house were valued at \$300K, does that seem surprising?
- Mathematically leverage the equation: calculus and optimization.

The line of best fit: the Least Squares criteria

The classical definition of the "best" line:

- Find the β_0 and β_1 that minimize

$$\sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2.$$

  Call the minimizers b_0 and b_1.
- In English, the "best line" minimizes the sum of the squares of the vertical distances from the data points to the line, and is called the *Least Squares Line*.
- Sometimes, we may fit a line on a transformed scale, then back-transform, which gives *best fitting* curves.
- The *Achilles heel* of Least Squares: it is extremely sensitive to points that are atypical in the x-direction (leverage).

- Call x the **predictor** variable and y the **response** variable.
- The fitted values are written as $\hat{y}_i$, and are calculated as $\hat{y}_i = b_0 + b_1 x_i$.
- The difference between y_i and $\hat{y}_i$, $y_i - \hat{y}_i$ is called the **residual**.
- We write the residual as e_i so that:

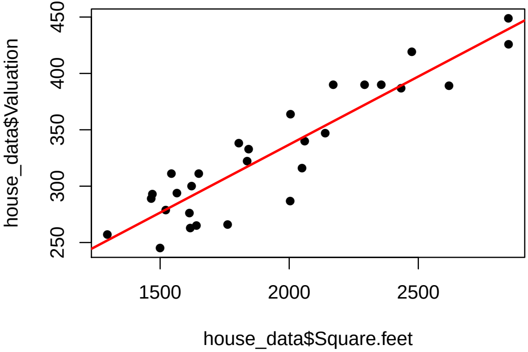
$$e_i = y_i - \hat{y}_i.$$

- The least squares estimates are given by:

$$b_1 = r\frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1\bar{x},$$

where r is the correlation between x and y.

Visualizing the regression



Interpreting the slope and intercept

- Don't forget the cardinal rule of data analysis: always, always plot the data. Interpreting the line doesn't make a lot of sense if it doesn't describe the data well.
- The fitted model: $\hat{y}_i = b_0 + b_1 x_i$.
  - Intercept: b_0: the expected value of y, when x = 0. It has the units of y.
  - Slope: b_1: the change in the expected value of y for every one unit change in x. Always understand the units on b_1. They are the units of y over the units of x.

At this point, focus on the regression prediction equation and coefficient interpretation:

$$\hat{y}_i = 95.49677 + 0.12070\, Square.feet.$$

Interpreting the slope and intercept

- The slope: each additional square foot adds approximately \$120 to the valuation (a little too causal).
- Safer: comparing two houses that differ in square footage by 100, we expect the larger one on average to be valued at an additional \$12,000. Maybe larger houses have more bathrooms, and maybe that is what's driving the increase in valuation.
- The intercept: that part of the value, that doesn't depend on the size of the house. Like a fixed cost. The value of the land perhaps?
- Also note that the intercept is a huge leap outside the range of the data, so that there could be plenty of extrapolation error.

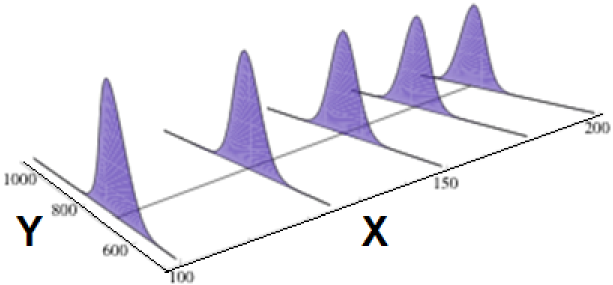
Regression assumptions

The Simple Regression Model (SRM) states that data is generated according to:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i are *independent and identically distributed (iid.)* and

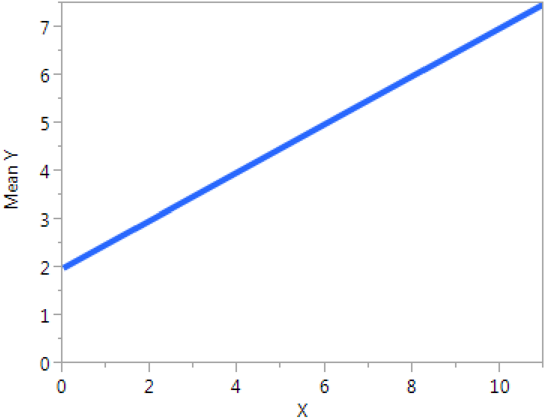
$$\epsilon_i \sim N(0, \sigma_\epsilon^2).$$



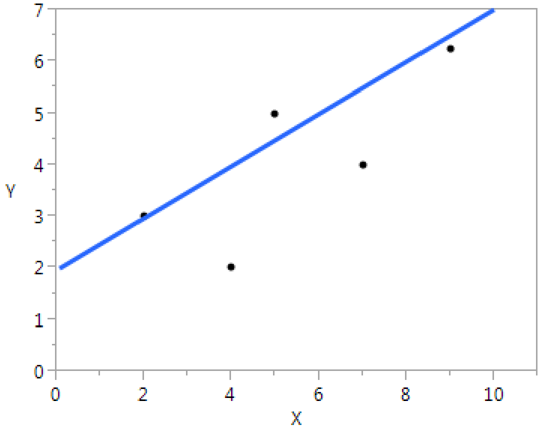
The regression story

There is a truth (but we don't know it), which is that there's a true regression line: $\beta_0 + \beta_1 x$.

$$E(Y|X) = 2 + 0.5X.$$

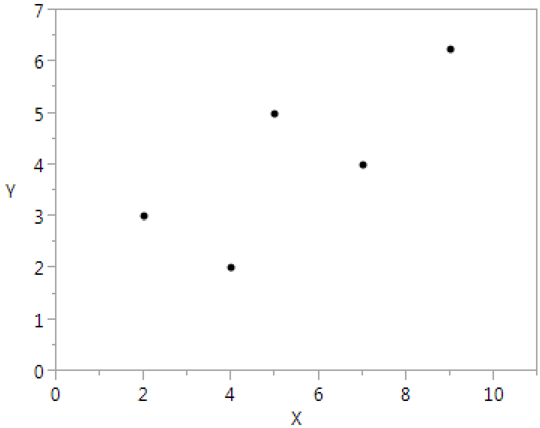


Data is generated according to this model



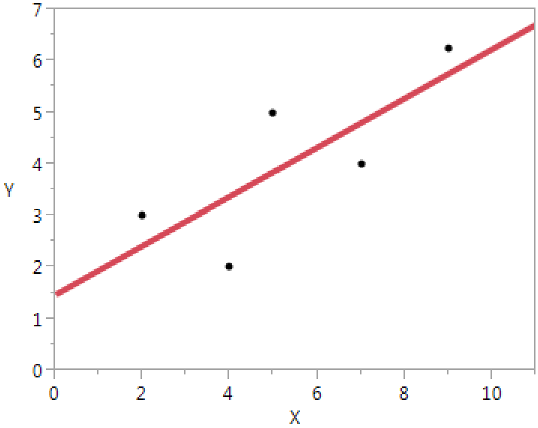
The data points have a normal distribution about this true line (the errors, ϵ_i).

The analyst's problem



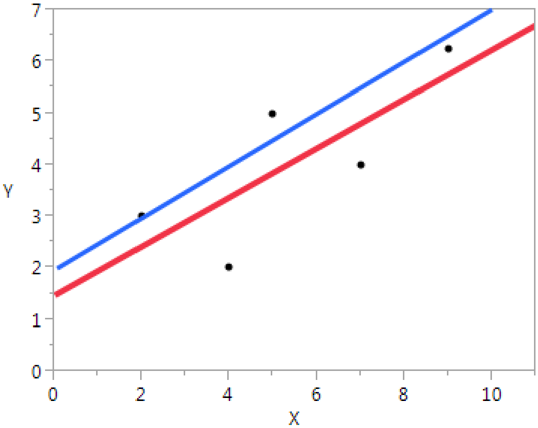
The analyst only has the data and tries to reconstruct the truth (*blue line*) using the method of least squares.

Overlaying the least squares line on the data



The least squares line is the red line. As humans, this is the best we can do: $\hat{y} = 1.48 + 0.48x$.

Hopefully least squares is close to the truth



We use the red line (least squares) as an approximation to the true blue line.

Model summaries: Root Mean Squared Error (RMSE)

- Fact: the sample mean of the residuals is always exactly zero.
- The standard deviation of the residuals, known as **R**oot **M**ean **S**quared **E**rror (RMSE):

$$RMSE = \sqrt{\frac{e_1^2 + e_2^2 + \cdots + e_n^2}{n-2}}.$$

- The (n - 2) in the denominator is there because we have estimated 2 parameters in the regression, the slope and intercept.
- RMSE is a measure of the residual variation, after modeling.
- It has the units of the y-variable.
- Low values of RMSE are good, and if you are choosing between models with the same outcome variable, then prefer models with the lower RMSE.

Using the RMSE in practice: prediction intervals

- RMSE is a critical regression summary.
- If the residuals are approximately normally distributed, then within the range of the data, an approximate 95% Prediction Interval for a **new** observation is:

$$\hat{y}_{new} \pm 2RMSE.$$

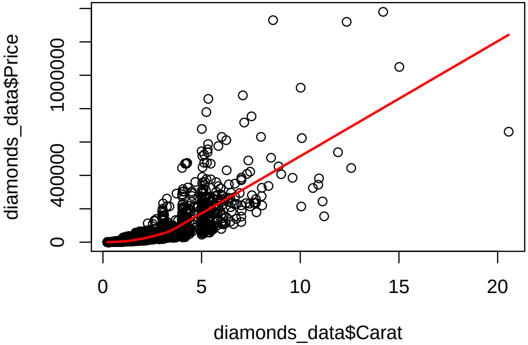
- For the housing data, the RMSE was 23.02.
- So 95% prediction intervals have a margin of error of approximately $46 thousand.

Model summaries: R^2

- Define R^2 as $(r)^2$, that is the sample correlation squared.
- Interpretation: the proportion of variability in y explained by the regression model.
- Facts about R^2:
    1. $0 \leq R^2 \leq 1$.
    2. An R^2 of 1 means perfect linear association.
    3. An R^2 of zero means no linear association.
    4. R^2 has no measurements units.
- All other things being equal, we prefer models with a higher R^2.
- But there is no magic R^2 value for the use of a model. The June 2019 used a dataset with an R^2 of only 1%!

The LOWESS smoother

- A data driven way of prospecting for curvature.
- There are many smoothers available, but `lowess` usually does a decent job.



Confidence and prediction bands for the regression: big picture

Recall that the model for data generation in regression is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

There are two types of interval associated with the fitted line:

1. Confidence bands for the regression.

$$y_i = \boxed{\beta_0 + \beta_1 x_i} + \epsilon_i.$$

2. Prediction intervals for a new observation.

$$y_i = \boxed{\boxed{\beta_0 + \beta_1 x_i} + \boxed{\epsilon_i}}.$$

Confidence bands for the regression

- The systematic (signal) part of the SRM:

$$E(Y|X) = \beta_0 + \beta_1 X.$$

- In English: what do you think the mean of Y is, for a given value of X, aka what's the height of the true regression line?
- Example: for houses with 2100 square feet, what is their expected valuation?
- We estimate this with the least squares regression line:

$$\hat{y} \approx b_0 + b_1 2100.$$

Confidence bands for the regression

- Provide a 95% CI for this expectation.
- To do this we need the standard error of

$$\hat{y}_i = b_0 + b_1 x_i.$$

- We can't just add the standard errors of b_0 and b_1 because this is a linear combination of random variables.
- But, if you did the math you would get:

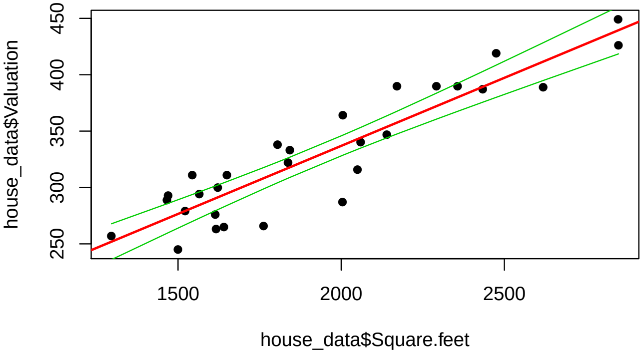
$$se(\hat{y}) = RMSE\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}.$$

Confidence bands for the regression

Three key observations from this formula:

1. This is a function of x_i.
2. As x_i moves away from $\bar{x}$ then this standard error increases (the statistical extrapolation penalty).
3. As n gets large the standard error goes to zero. The confidence bands will collapse and capture the true regression line.

Confidence bands for the regression



The green bands are the 95% confidence bands for the regression, and capture the uncertainty in the estimate of the regression line itself.

Prediction intervals for a **new** observation

- If we were to draw one new observation, then based on the fitted regression line, where do you think its y-value will be?
- Notice, that this is not a question about the average value of y, but the single realized new value.
- Example: I have a house with 2100 square feet. Provide a 95% PI for its valuation.

Prediction intervals for a **new** observation

We are trying to estimate the quantity:

$$y_{new} = \beta_0 + \beta_1 x_{new} + \epsilon_{new}.$$

To construct the interval we need to estimate this quantity and obtain its standard error:

$$\hat{y}_{new} = b_0 + b_1 x_{new},$$

$$se(\hat{y}_{new}) = RMSE\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{(n-1)s_x^2}}.$$

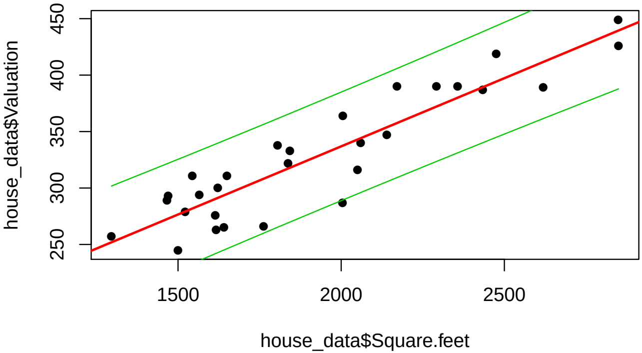
Notice that the forecast is the same as before $(b_0 + b_1 x_{new})$ but the standard error has changed to incorporate the additional ϵ term.

Prediction intervals for the regression

Three key observations from this formula:

1. This is a function of x_i.
2. As x_i moves away from $\bar{x}$ then this standard error increases (the statistical extrapolation penalty).
3. As n gets large the standard error goes to RMSE $\approx \sigma_\epsilon$, **not** zero. The 95% prediction intervals will become parallel to the true regression line at a distance of approximately $+/$-2RMSE.

Prediction intervals



The green bands are the 95% prediction intervals, and capture the precision of the predictions for new observations.

Summary

Topics covered:

- The Simple Regression Model (SRM).
- Fitting via Least Squares.
- Interpretation of regression coefficients.
- Assumptions and residual checking.
- Model summaries, R^2 and RMSE.
- Confidence bands and prediction intervals for the regression line.