Extra Considerations

David Gerard

2018-12-07

Learning Objectives

- Different interval estimates at different levels of the explanatory variable.
- Extrapolation vs interpolation
- Correlation
- R^2

Model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- Y_i: distance from earth of nebula i
- X_i: recession velocity of nebula i
- β_0: The intercept of the mean line ("regression line")
  - Mean when $X_i = 0$

Model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- Y_i: distance from earth of nebula i
- X_i: recession velocity of nebula i
- β_0: The intercept of the mean line ("regression line")
  - Mean when $X_i = 0$
- β_1: Slope of the regression line.
  - Difference in mean distance between two nebula when they differ by only 1 velocity unit.

Model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- Y_i: distance from earth of nebula i
- X_i: recession velocity of nebula i
- β_0: The intercept of the mean line ("regression line")
  - Mean when $X_i = 0$
- β_1: Slope of the regression line.
  - Difference in mean distance between two nebula when they differ by only 1 velocity unit.
- $\beta_0 + \beta_1 X_i$: the mean distance at velocity X_i

Model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- Y_i: distance from earth of nebula i
- X_i: recession velocity of nebula i
- β_0: The intercept of the mean line ("regression line")
  - Mean when $X_i = 0$
- β_1: Slope of the regression line.
  - Difference in mean distance between two nebula when they differ by only 1 velocity unit.
- $\beta_0 + \beta_1 X_i$: the mean distance at velocity X_i
- ϵ_i: Individual noise with mean 0 and variance σ^2. Ideally normally distributed.

Various Intervals

Pointwise confidence intervals

- Suppose we want to estimate the mean at a single value of X_0.
- Parameter: $\beta_0 + \beta_1 X_0$
- Point Estimate: $\hat{\beta}_0 + \hat{\beta}_1 X_0$
- Confidence interval: estimate + multiplier * standard error

Pointwise confidence intervals

- You can show that the standard error of $\hat{\beta}_0 + \hat{\beta}_1 X_0$ is

$$\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s_X^2}}$$

- You can also show that

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 X_0 - (\beta_0 + \beta_1 X_0)}{SE(\hat{\beta}_0 + \hat{\beta}_1 X_0)} \sim t_{n-2}$$

- You an use this t-ratio in the usual ways to run hypothesis tests and get confidence intervals.

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{n-2}(0.975) SE(\hat{\beta}_0 + \hat{\beta}_1 X_0)$$

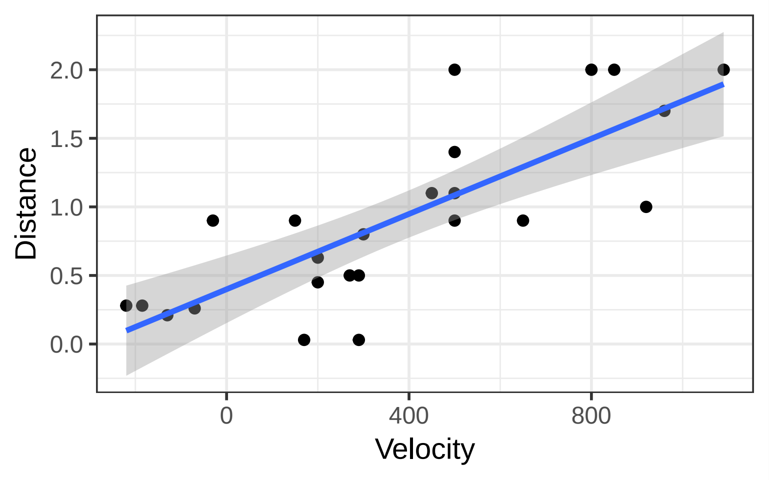
Point-wise confidence intervals in R

```r
library(Sleuth3)
data("case0701")
lmout <- lm(Distance ~ Velocity, data = case0701)
predict(lmout, newdata = data.frame(Velocity = 100),
        interval = "confidence")
```

```
##      fit    lwr    upr
## 1 0.5364 0.3216 0.7512
```

Plotting pointwise confidence intervals in R

```
library(ggplot2)
qplot(Velocity, Distance, data = case0701, geom = "point") +
  geom_smooth(method = "lm")
```



Simultaneous Confidence Bands

- Sometimes you want to ask "Where is the regression line?"
- You can capture the **entire regression line** with 95% confidence with

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm \sqrt{2F_{2,n-2}(0.95)}SE(\hat{\beta}_0 + \hat{\beta}_1 X_0)$$

- $F_{2,n-2}(0.95)$ is the 95th percentile of an F distribution with 2 numerator degrees of freedom and $n-2$ denominator degrees of freedom.
- No easy way to get these bands in R without a third package.
- The bands in `qplot()` are **pointwise** confidence intervals, **not** simultaneous confidence bands

Prediction intervals

- Sometimes, you want to get likely values for future observations at a given value of X_0
- Answers question "what are likely distances of a nebula at a given velocity?"
- This is **different** from "what are likely mean distances at a given velocity?"

Prediction intervals

- Predict a future observation with its estimated mean.
- Variability in prediction consists of two components.

$$
\begin{aligned}
Y - Pred(Y|X_0) &= Y - (\hat{\beta}_0 + \hat{\beta}_1 X_0) \\
&= Y - (\beta_0 + \beta_1 X_0) + [(\beta_0 + \beta_1 X_0) - (\hat{\beta}_0 + \hat{\beta}_1 X_0)]
\end{aligned}
$$

- Variance of first term is σ^2
- Variance of second term is $SD(\hat{\beta}_0 + \hat{\beta}_1 X_0)$
- Variance of prediction is sum of these two variances.

Prediction intervals

- Variance of $Y - Pred(Y|X_0) = \sigma^2 + SD(\hat{\beta}_0 + \hat{\beta}_1 X_0)$
- Standard error of prediction is $\sqrt{\hat{\sigma}^2 + SE(\hat{\beta}_0 + \hat{\beta}_1 X_0)}$
- Prediction interval is

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{n-2}(0.975)\sqrt{\hat{\sigma}^2 + SE(\hat{\beta}_0 + \hat{\beta}_1 X_0)}$$

Prediction intervals

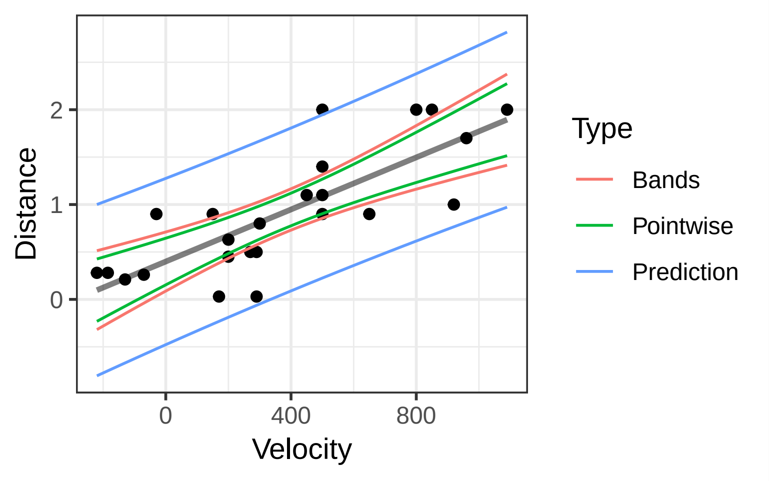
- For prediction intervals, **the central limit theorem does not save us**.
- Prediction intervals are **very** sensitive to violations in normality.
- This is because we are trying to account for the variability in **a single observation**.

Prediction intervals in R

```r
predict(lmout, newdata = data.frame(Velocity = 100),
        interval = "prediction")
```

```
##       fit     lwr   upr
## 1 0.5364 -0.3318 1.405
```

Comparison of Intervals

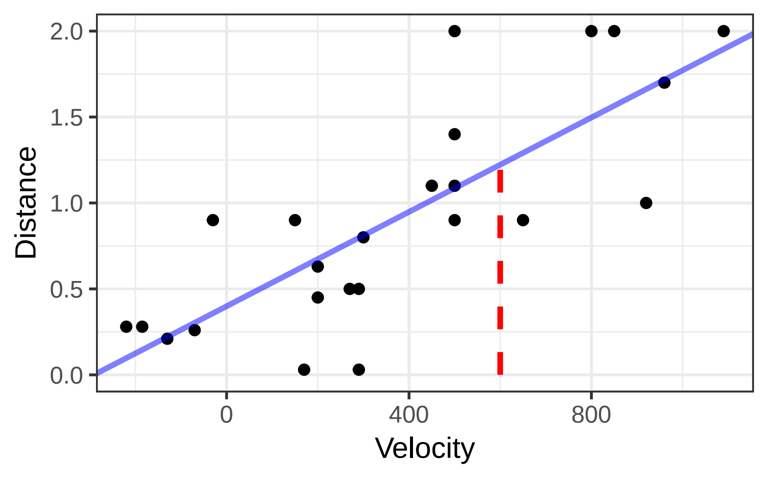


Extrapolation vs Interpolation

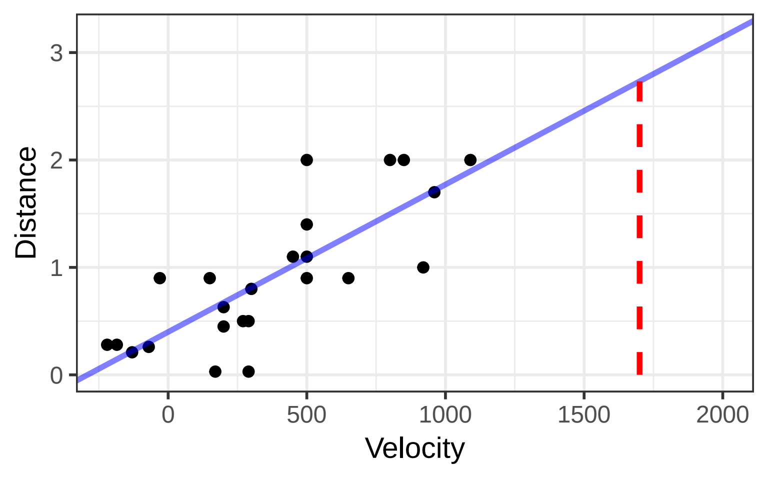
Definitions

- **Interpolation**: Making estimates/predictions within the range of the data.
- **Extapolation**: Making estimates/predictions outside the range of the data.
- Interpolation is good. Extrapolation is bad.

Interpolation



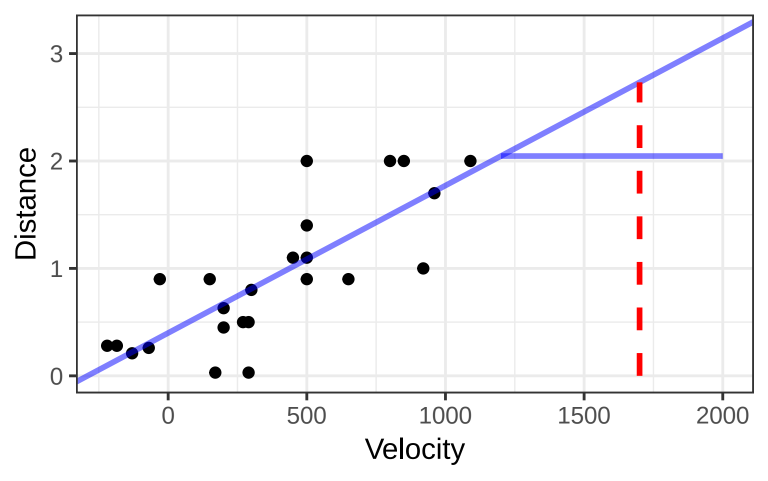
Extrapolation



Why is extrapolation bad?

1. Not sure if the linear relationship is the same outside the range of the data (because we don't have data there to see the relationship).
2. Not sure if the variability is the same outside the range of the data (because we don't have data there to see the variability).

Changing Relationship



Correlation

Correlation

- Sample correlation is a measure of **linear** association.

$$
\begin{aligned}
r_{XY} &= \frac{Average\left((X_i - \bar{X})(Y_i - \bar{Y})\right)}{s_X s_Y} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{s_X s_Y}
\end{aligned}
$$

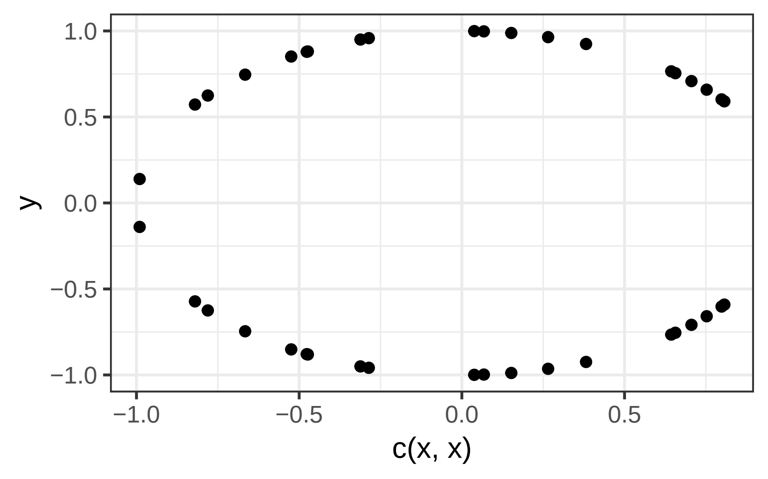
Correlation Properties

- No units.
- Always between -1 and 1.
- Closer to 0 means less **linear** association.
- Closer to 1 or -1 means stronger **linear** association.
- Correlation = -1 or 1 if and only if all points lie **exactly** on a straight line.
- Useful as a summary statistic. Not usually useful for **inference**

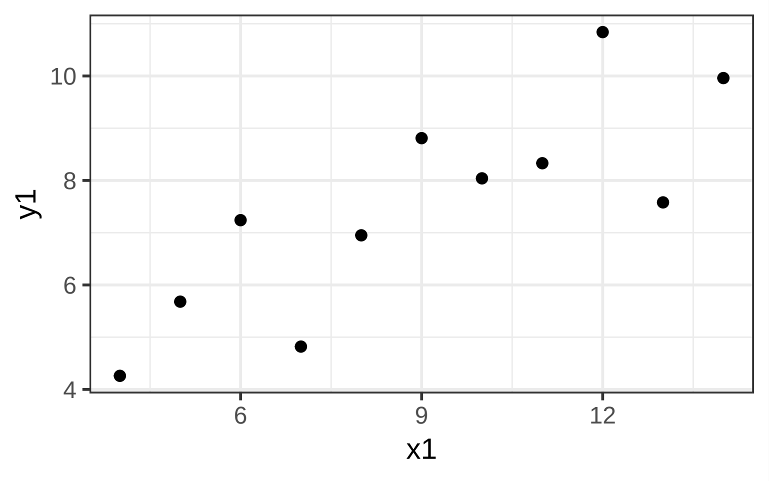
Correlation

- Correlation can be misleading so **always** plot data

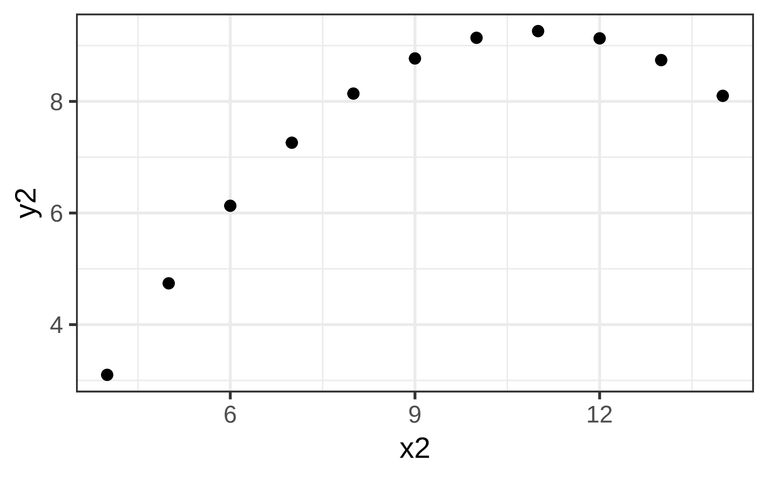
Correlation of 0, but Very Associated



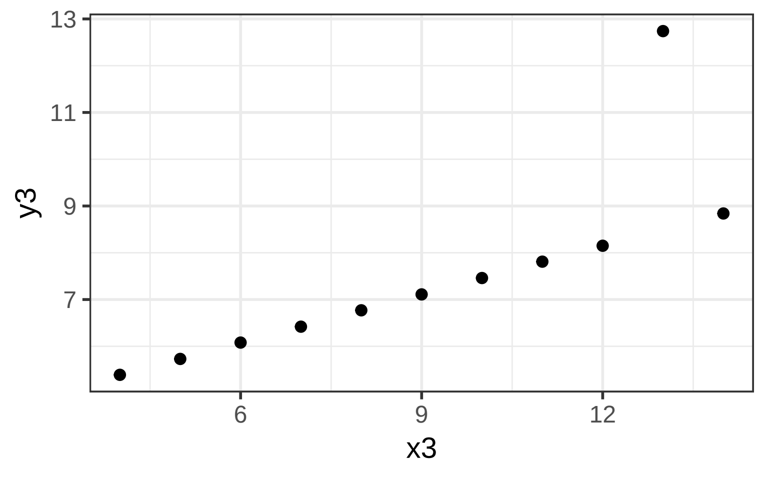
All these datasets have a correlation of 0.81



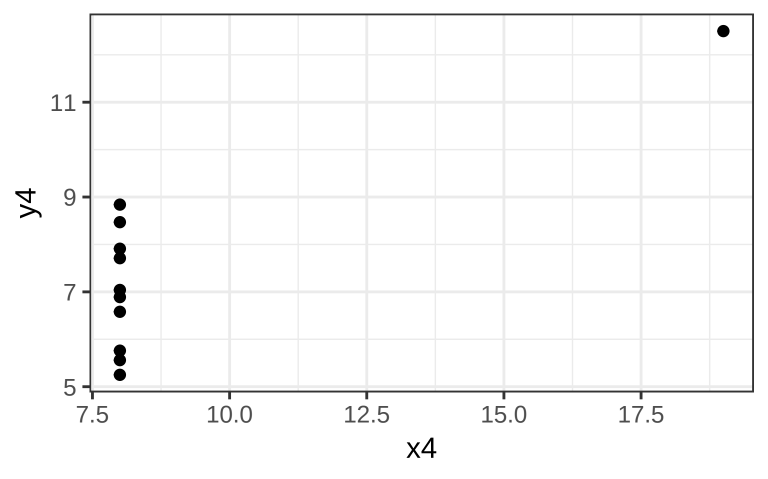
All these datasets have a correlation of 0.81



All these datasets have a correlation of 0.81



All these datasets have a correlation of 0.81

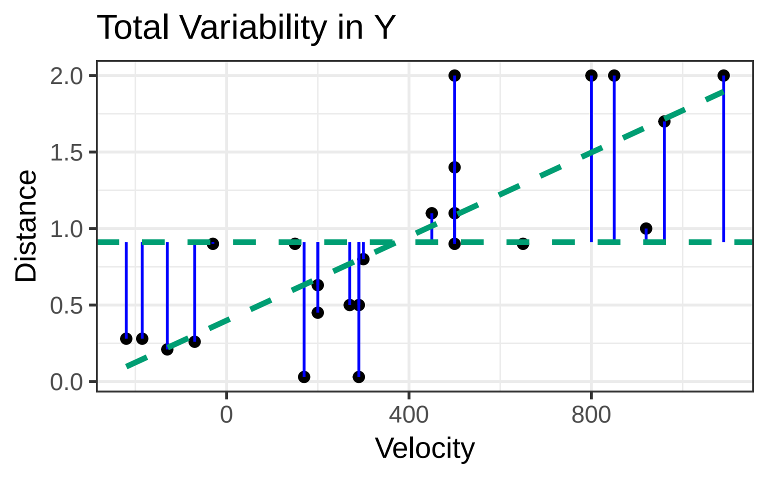


Correlation Intuition

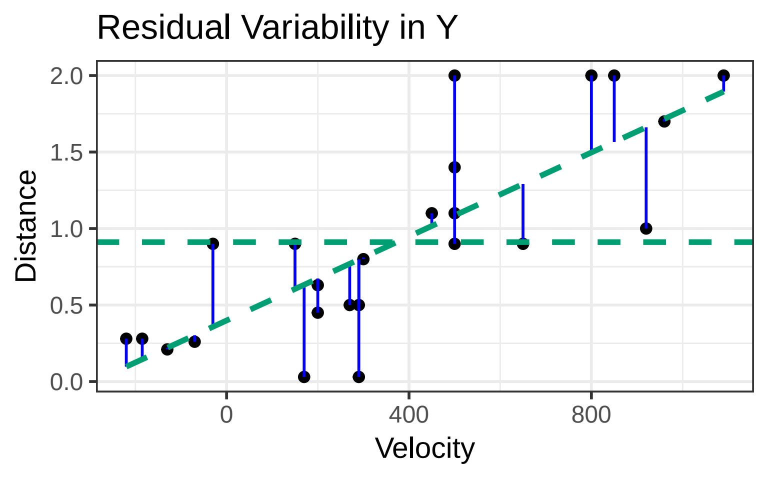
http://guessthecorrelation.com/

R^2 (Section 8.6)

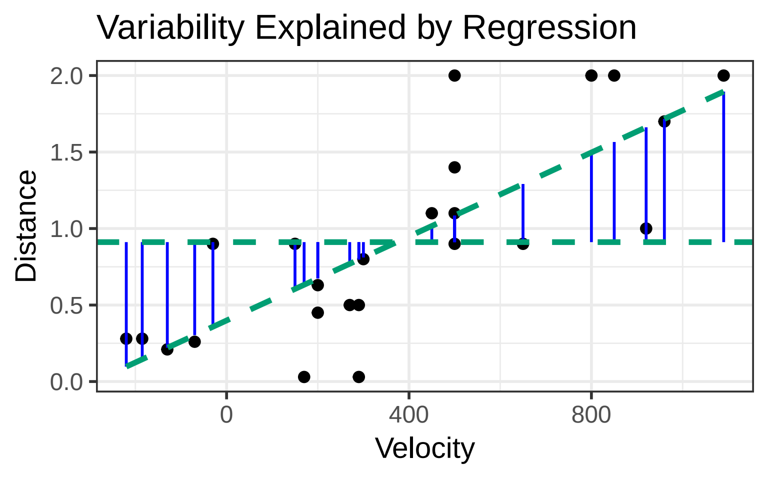
Total Variability in Y



Residual Variability in Y



Variability in Y explained by regression line



R^2

$$R^2 = \frac{\text{Total Sum of Squares} - \text{Residual Sum of Squares}}{\text{Total Sum of Square}}$$
$$= \frac{\text{Extra Sum of Squares}}{\text{Total Sum of Square}}$$

R^2 Properties

- Proportion of variation explained by the regression line.
- Close to 0 means weak linear relationship.
- Close to 1 means strong linear relationship.
- In physics, $R^2 = 0.99$ is good, $R^2 = 0.9$ is bad.
- In social science and humanities, $R^2 = 0.25 - 0.5$ is really good.
- In biology, you want R^2's somewhere between those two.

R^2 and Correlation

- The R^2 is **exactly** the correlation between X and Y squared.
- Useful as a summary statistic, not useful for inference.
- **Cannot** use it to evaluate the fit of a linear regression line (same problems as correlation).