Module 2
The Simple Regression Model

The Signal-Noise Decomposition

One of the ideal setups in Statistics occurs when data

\[ y_1, \ldots, y_n \]

can be treated as sequence of independent draws from a normal distribution with mean \( \mu_y \) and standard deviation \( \sigma_y \).

This statistical model for such data is denoted by

\[ y_1, \ldots, y_n \sim iid \sim N(\mu_y, \sigma_y^2) \]

where \( iid \) stands for “independent and identically distributed”.

Example

In Stat 603 we saw that the \( iid \) normal model was reasonable for the 1992-1993 daily returns on GM (see GM92.jmp).

There, based on \( n = 507 \) observations, we estimated

\[ \mu_y \text{ by } \bar{y} = .00158, \text{ and } \sigma_y \text{ by } s_y = .0202 \]

Somewhat more suggestively, this model can also be written as

\[ y_i = \mu_y + \epsilon_i, \quad i = 1, \ldots, n \]
\[ \epsilon_1, \ldots, \epsilon_n \sim iid \sim N(0, \sigma^2_\epsilon) \]

Notice how the data generating process has two components\(^1\):

1) “signal”: a fixed level \( \mu_y \)

2) “noise”: \( \epsilon_1, \ldots, \epsilon_n \) iid mean 0 normal deviations

Although the \( iid \) normal model above is not always appropriate, it is a special case of a broadly applicable model formulation

\[ y_i = \text{signal}_i + \epsilon_i, \quad i = 1, \ldots, n \]
\[ \epsilon_1, \ldots, \epsilon_n \sim iid \sim N(0, \sigma^2_\epsilon) \]

Again, the data generating process has two components:

1) the signal: \( signal_i \)

2) the noise: \( \epsilon_1, \ldots, \epsilon_n \) iid mean 0 normal deviations

Note the three main properties of the noise \( \epsilon_1, \ldots, \epsilon_n \)

a) independence

b) equal variance \( \sigma^2_\epsilon \)

c) normally distributed

\(^1\) The terminology “signal” and “noise” originated in electrical engineering. The methods we are studying can also be used to improve the reception of a TV or radio station. The goal of engineers is to transmit a clear signal from the station, one free of noise. For us, signal is an underlying structure that we seek to separate from random noise.
Although the decomposition
\[ y_i = \text{signal}_i + \epsilon_i \]
is not explicitly observed, a general strategy for finding such a model is based on finding a decomposition of the form
\[ y_i = \hat{y}_i + e_i, \quad i = 1, \ldots, n \]
where \( \hat{y}_i \) estimates \( \text{signal}_i \)
and \( e_i = y_i - \hat{y}_i \) estimates noise \( \epsilon_1, \ldots, \epsilon_n \).

Support for a particular form for \( \text{signal}_i \) is obtained when \( e_1, \ldots, e_n \) manifest iid normal behavior.

**Jargon**

\( \hat{y}_1, \ldots, \hat{y}_n \) are called the *fitted* or *predicted values*

\( e_1, \ldots, e_n \) are called the *residuals*

As we shall now see, regression analysis falls exactly into this framework.

**The Simple Regression Model (SRM)**

Under this idealized statistical model, the data 
\((x_1, y_1), \ldots, (x_n, y_n)\)
are a realization of
\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n \]
where \( \epsilon_1, \ldots, \epsilon_n \) iid \( \sim \mathcal{N}(0, \sigma^2) \)

Pictorially:

As a decomposition of the data into signal & noise, the signal here is

and the noise here is

People also sometimes refer to \( \epsilon_1, \ldots, \epsilon_n \) as the “errors”.\(^2\)

\(^2\) You can also think of these errors as coming from all of the other factors that influence the response aside from the one that we have chosen to highlight in the simple regression.
Think of the SRM as a hypothetical process that could have generated the data.

Example
To get a feel for how the SRM generates data, the file Utopia.jmp contains a simulation of pairs

\[(x_1, y_1), \ldots, (x_n, y_n)\]

from a SRM with \(\beta_0 = 7, \ \beta_1 = .5, \ \text{and } \sigma_e = 1\)

What are the interpretations of \(\beta_0 + \beta_1 x, \beta_0, \beta_1\) and \(\sigma_e\) in the SRM?

\(\beta_0, \beta_1\) and \(\sigma_e\) are the (usually) unknown parameters of the SRM. An objective of regression is to estimate them.

---

Typical Regression Situation
The general course of a regression analysis includes these steps:

1. Figure out for your problem if it makes sense to think of one variable as a predictor, and one as a response.
2. Observe pairs of data, \((x_1, y_1), \ldots, (x_n, y_n)\)
3. Plot the data!
4. If necessary, transform the data to obtain linear association
5. Suspect (or hope) SRM assumptions are justified
6. Estimate the “true” regression line

\[y = \beta_0 + \beta_1 x\]

by the LS regression line

\[\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x\]

where \(\hat{\beta}_0\) and \(\hat{\beta}_1\) to denote \(b_0\) and \(b_1\) from Module 1.

WARNING! The true regression line and the LS regression line are different. DON'T CONFUSE THEM!

Pictorially

\[\hat{\beta}_0, \hat{\beta}_1\] are often referred to as the least squares (LS) estimates of \(\beta_0\) and \(\beta_1\).
The Fitted Values and the Residuals

The LS regression line decomposes the data into two parts

\[ y_i = \hat{y}_i + e_i \]

where

\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{and} \quad e_i = y_i - \hat{y}_i \]

Pictorially

Jargon (again)

\( \hat{y}_1, \ldots, \hat{y}_n \) are called the fitted or predicted values

\( e_1, \ldots, e_n \) are called the residuals

The following page shows the fitted values and the residuals for the Module 1 diamond regression.\(^4\)

\[^4\] After executing the Fit Line subcommand, JMP will store the fitted values and residuals in the data table by right clicking next to “--Linear Fit” and selecting Save Predicteds and Save Residuals form the Pop-up menu.

Does the decomposition \( y_i = \hat{y}_i + e_i \) hold here?
Root Mean Squared Error (RMSE) – An Estimate of $\sigma_e$

Looking at more of the output from the diamond regression, a key quantity of interest is the

Root Mean Square Error (RMSE) $= 31.84$

$RMSE^2$ is the “average” squared deviation between the data and the LS regression line (i.e. the variance of the residuals).

We divide by $(n - 2)$ instead of $n$ to compensate for the fact that the LS line obtains smaller sum of squared deviations than the true regression line$^5$.

How does the formula for $RMSE$ compare to the formula for $s_y$, the sample standard deviation of $y$?

$RMSE$ measures the dispersion of the residuals around the LS regression line. Why is this value important in the regression?

If the SRM holds, then approximately

of the data will lie within one $RMSE$ of the LS line

of the data will lie within two $RMSE$ of the LS line

$^5$ The quantity $(n - 2)$ here is sometimes called the degrees of freedom (df) and is often used in regression calculations.
Model Checking

Any conclusions drawn from a regression analysis depend on the assumption that the SRM is appropriate.

Good statistical practice entails using the data to make sure there are no gross violations of the SRM.

What to look for:

1) Is the relationship between \( x \) and \( y \) linear?

2) Are there outliers or influential values that distort the model fit?

3) Do the residuals manifest iid normal behavior? (i.e., independent, constant variance, normal)

Three crucial model checks:

1. A scatterplot of \( y \) vs \( x \) should reveal

2. A scatterplot of the Residuals vs \( x \) should appear

3. A histogram of the residuals should appear and a normal quantile plot of the residuals should appear

Example: Checking the Diamond Regression

The scatterplot of Price vs Weight (p 2-9) reveals

The scatterplot\(^6\) of Residuals vs Weight reveals

The histogram and normal quantile plot of the residuals shows

---

\(^6\) After executing the Fit Line subcommand, right click on the triangle next to “---Linear Fit” and select Plot Residuals from the Pop-up menu to obtain this plot.
Anomalies to Look For

Nonlinearity

Can be revealed by the $y$ vs $x$ scatterplot or by the Residuals vs $x$ scatterplot

Recall the display.jmp data

Remedy: Transform $y$ and/or $x$.

Autocorrelated Residuals

The file cellular.jmp contains the number of subscribers to cellphone service in the US every six months from the end of 1984 to the end of 1995.

The data is a time series $y_1, \ldots, y_n$ where $y_i$ is the number of subscribers at time period $t$.

A scatterplot of $y$ vs $t$ (i.e. a time series plot of $y$) shows nonlinear growth in the number of subscribers.

By trial and error\(^7\), one discovers that the transformation $y^* = y^{1/4}$ yields what appears to be an ideal linear relationship.

Thus one might consider fitting a trend model of the form

$$y_t^{1/4} = \beta_0 + \beta_1 t + \epsilon_t, \quad t = 1, \ldots, n$$

In this special case of the SRM, $t$ plays the role of $x$.

\(^7\) As described in Lecture 1 of BAR, p 29-38.
At first glance the regression of $y^{1/4}$ on $t$ appears to be wonderful.

However, the scatterplot of residuals vs $t$ reveals a serious problem.

What SRM assumption has been violated?

Such meandering residuals are often called autocorrelated because $e_{t-1}$ and $e_t$ appear correlated.

Expanding Residuals

The file cleaning1.jmp contains the number of crews (Crews) and the number of rooms cleaned (RoomsClean) for 53 teams of building maintenance workers.

Which assumption of the SRM is violated here?

This violation has only a minor effect on the estimation of $\beta_0$ and $\beta_1$. However, it does affect the prediction statements to be discussed in Module 3.

Remedy: Transform $y$ or use weighted least squares (p 57-60) instead of least squares.
Outliers and Influential Points

Main idea: outliers are unusual points. They should always be investigated. If warranted, they should be excluded.

The file direct.jmp contains the level of sales (\$1000’s) and the number of direct mail recipients (1000’s) for 10 different mailings of a catalog.

Each catalog costs $1.50 and the company would like to assess its marginal profit increase per catalog.

Investigation of the unusual point above reveals that that mailing coincided with a large inventory sale.

Repeating the regression with the point excluded yields

The LS line here is \( \text{Sales} = 5.78 + 1.98 \text{Direct} \) with RMSE = 8.8

What changed?

Should the outlier be excluded?

What is the company’s estimated marginal profit per catalog?
Another Outlier Example
The file phila.jmp contains the average prices of houses sold in the prior year and crime rates for 110 Pennsylvania communities in and near Philadelphia in April 1996.

To gauge the relationship between house prices and crime rates, one might consider the following regression:

$$\text{House Price} = 176629 - 577 \text{ Crime Rate}$$

with RMSE = 84325. Interpretation?

The unusual point is

Repeating the regression with the unusual point excluded yields

$$\text{House Price} = 225234 - 2289 \text{ Crime Rate}$$

with RMSE = 78861. How does this fit change the implications of the previous model?

Note how one point can drastically influence a regression. Should this point be excluded?

How does the unusual point affect the fit here, compared to the effect of the outlier in the previous direct mail regression?

---

8 Point labels are very helpful when it comes to identifying outliers. The default point label is the row number in the JMP data set. You can assign a variable to be the label as well.
Don't forget to plot the data!

Before fitting a regression, it is crucial to first plot the data.

Example: Which of the following four data sets seems compatible with the SRM assumptions?

![Plot of data set 1](image1)

![Plot of data set 2](image2)

![Plot of data set 3](image3)

![Plot of data set 4](image4)

Which of the previous scatter plots yields the following regression output?

### Bivariate Fit of Y By X

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>7.5</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>12.5</td>
<td>20</td>
</tr>
</tbody>
</table>

**Linear Fit**

\[ Y = 3.000909 + 0.5000909 X \]

**Summary of Fit**

- **RSquare**: 0.666542
- **RSquare Adj**: 0.629492
- **Root Mean Square Error**: 1.236603
- **Mean of Response**: 7.500909
- **Observations (or Sum Wgts)**: 11

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>27.510001</td>
<td>27.5100</td>
<td>17.9899</td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>Error</td>
<td>9</td>
<td>13.762690</td>
<td>1.5292</td>
<td></td>
<td>0.0022</td>
</tr>
<tr>
<td>C. Total</td>
<td>10</td>
<td>41.272691</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | Prob>|t| |
|---|---|---|---|---|
| Intercept | 3.0000909 | 1.124747 | 2.67 | 0.0257 |
| X | 0.5000909 | 0.117906 | 4.24 | 0.0022 |
Take-Away Summary
The simple regression model (SRM) is the basis for inference from regression with one predictor. In this model, the observed data

$$(x_1, y_1), \ldots, (x_n, y_n)$$

are assumed to be a realization of a “signal+noise” data generating process that ideally has the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\varepsilon_1, \ldots, \varepsilon_n \text{iid } \sim \mathcal{N}(0, \sigma^2)$$

Important diagnostics to keep in mind are plots that check for
- Outliers
- Linearity
- Independence when the data are ordered
  (particularly when the data are a time series)
- Equal variance
- Normality

Next Module
The SRM is the basis for confidence intervals, prediction intervals, and hypothesis tests.