# STAT22000 Autumn 2013 Lecture 5

Yibi Huang

October 9, 2013

2.3 Least-Squares Regression

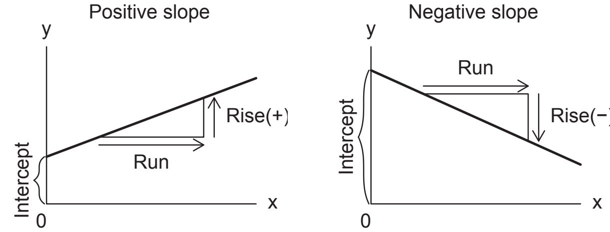


## Math Review: Equation of a Straight Line

The equation of a straight line is of the form

$$y = \text{intercept} + \text{slope} \times x$$

.




$$\text{Slope} = \frac{\text{Rise}}{\text{Run}}$$

In a regression problem, $x$ is the explanatory variable, and $y$ is the response variable.



## Explanatory and Response Variables

In a regression problem, one variable is predicted or explained based on one or several other variables.

- The variable to be predicted is called the **response variable**, or just the **response**.
- The variable(s) to predict or to explain the variation in the response is called the **explanatory variable(s)**

---

**Remark**: Some books call the response the **dependent variable**, and the explanatory variable the **independent variable**. We don't use these terms because "dependence" and "independence" have other meanings in statistics.

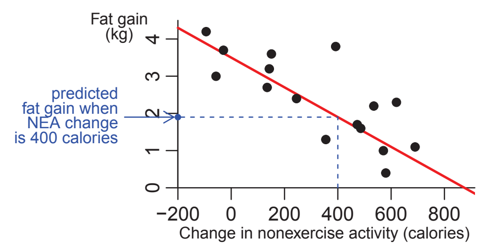


## Example 2.12 Fidgeting and Fat Gain (p.109)

| NEA change (cal) | Fat gain (kg) |
|---|---|
| −94 | 4.2 |
| −57 | 3.0 |
| −29 | 3.7 |
| 135 | 2.7 |
| 143 | 3.2 |
| 151 | 3.6 |
| 245 | 2.4 |
| 355 | 1.3 |
| 392 | 3.8 |
| 473 | 1.7 |
| 486 | 1.6 |
| 535 | 2.2 |
| 571 | 1.0 |
| 580 | 0.4 |
| 620 | 2.3 |
| 690 | 1.1 |

Say we predict fat gain ($y$) from NEA change ($x$) using an (arbitrary) straight line

$$y = 3.5 - 0.004x$$




When NEA increases by 400 calories ($x = 400$), the predicted fat gain is

$$y = 3.5 - 0.004 \times 400 = 1.9kg$$

How good is this prediction?



## Predicted Values and Residuals (1)

We can assess the goodness of fit of a line by comparing the predicted $y$'s with the observed $y$'s.

For example, say we again use the line

$$y = 3.5 - 0.004x.$$

For an observation $(x_i, y_i)$, the **predicted value** for $y$, denoted as $\widehat{y}_i$, is

$$\widehat{y}_i = 3.5 - 0.004x_i,$$

and the **residual** (or **prediction error**) $e_i$ is the difference of the observed $y_i$ and the predicted $\widehat{y}_i$

$$e_i = y_i - \widehat{y}_i = y_i - (3.5 - 0.004x_i)$$

See the predicted values and residuals for NEA and fat gain data using the line $y = 3.5 - 0.004x$ on the next slide.



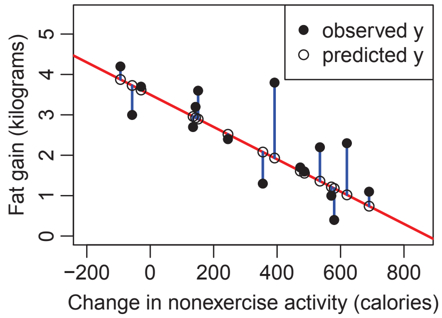
## Predicted Values and Residuals (2)

| NEA change $x_i$ (cal) | Fat gain $y_i$ (kg) | Predicted fat gain $\widehat{y}_i = 3.5 - 0.004x_i$ (kg) | Residual $e_i = y_i - \widehat{y}_i$ (kg) |
|---|---|---|---|
| −94 | 4.2 | $3.5 - 0.004 \times 4.2 = 3.88$ | $4.2 - 3.88 = 0.32$ |
| −57 | 3.0 | $3.5 - 0.004 \times 3.0 = 3.73$ | $3.0 - 3.73 = -0.73$ |
| −29 | 3.7 | $3.5 - 0.004 \times 3.7 = 3.62$ | $3.7 - 3.62 = 0.08$ |
| 135 | 2.7 | $3.5 - 0.004 \times 2.7 = 2.96$ | $2.7 - 2.96 = -0.26$ |
| 143 | 3.2 | $3.5 - 0.004 \times 3.2 = 2.93$ | $3.2 - 2.93 = 0.27$ |
| 151 | 3.6 | $3.5 - 0.004 \times 3.6 = 2.90$ | $3.6 - 2.90 = 0.70$ |
| 245 | 2.4 | $3.5 - 0.004 \times 2.4 = 2.52$ | $2.4 - 2.52 = -0.12$ |
| 355 | 1.3 | $3.5 - 0.004 \times 1.3 = 2.08$ | $1.3 - 2.08 = -0.78$ |
| 392 | 3.8 | $3.5 - 0.004 \times 3.8 = 1.93$ | $3.8 - 1.93 = 1.87$ |
| 473 | 1.7 | $3.5 - 0.004 \times 1.7 = 1.61$ | $1.7 - 1.61 = 0.09$ |
| 486 | 1.6 | $3.5 - 0.004 \times 1.6 = 1.56$ | $1.6 - 1.56 = 0.04$ |
| 535 | 2.2 | $3.5 - 0.004 \times 2.2 = 1.36$ | $2.2 - 1.36 = 0.84$ |
| 571 | 1.0 | $3.5 - 0.004 \times 1.0 = 1.22$ | $1.0 - 1.22 = -0.22$ |
| 580 | 0.4 | $3.5 - 0.004 \times 0.4 = 1.18$ | $0.4 - 1.18 = -0.78$ |
| 620 | 2.3 | $3.5 - 0.004 \times 2.3 = 1.02$ | $2.3 - 1.02 = 1.28$ |
| 690 | 1.1 | $3.5 - 0.004 \times 1.1 = 0.74$ | $1.1 - 0.74 = 0.36$ |

The residuals can tell us how good our prediction is.
E.g., the SD for these 16 residuals is $\approx$ 0.73kg, we can then expect that our prediction might be off by 0.73kg "on average".

## Predicted Values and Residuals on the Scatter Plot

- For an observed point $(x_i, y_i)$, the **predicted** $\widehat{y}_i$ is the vertical projection of the point to the line.
- The **residuals** are the signed distance from the observed points to the predicted points (the blue vertical segments, positive for points above the line, negative for below.)






## The Least Square Line

In general, we want to find a straight line $y = a + bx$ with small residuals

$$e_i = y_i - \widehat{y}_i = y_i - (a + bx_i).$$

However, it is impossible to minimize all residuals simultaneously (unless all points lie on a straight line). If one residual is reduced, often some other residuals will increase in size. We can only try to minimize the overall error. The **least squares regression line** of $y$ on $x$ is the line $y = a + bx$ that minimizes the sum of squared errors:

$$\sum_{i=1}^{n}(\text{residuals})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

and the line has slope

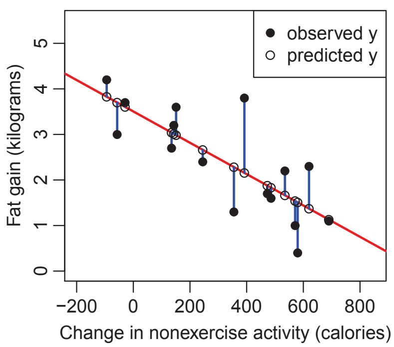
$$\text{slope} = \widehat{b} = r\frac{s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{and}$$

$$\text{intercept} = \widehat{a} = \overline{y} - \text{slope} \cdot \overline{x}$$



## The Least Square Line (2)

Graphically, the **least-square regression line** is the line that minimizes the *sum of squared vertical distances* from the points to the line, i.e., $\sum_{i=1}^{n}(\text{lengths of the blue vertical segments})^2$.




Note it is NOT minimizing the shortest distances but the vertical distances, because the shortest distances are not residuals but the vertical distances are.



## Example 2.12 Fidgeting and Fat Gain (p.109)

| | NEA change ($x$) | Fat gain ($y$) |
|---|---|---|
| mean | 324.75 | 2.3875 |
| SD | 257.66 | 1.1389 |

, $r = -0.7786$

The slope and intercept of the least square regression line to predict fat gain ($y$) from NEA change ($x$) are

$$\text{slope} = r\frac{s_y}{s_x} = -0.7786 \times \frac{1.1389}{257.66} \approx -0.00344$$

$$\begin{aligned}\text{intercept} &= \overline{y} - \text{slope} \times \overline{x} \\ &= 2.3875 - (-0.00344) \times 324.75 \approx 3.504\end{aligned}$$

So the least square regression line is $y = 3.504 - 0.00344x$, i.e.,

$$\text{predicted fat gain} = 3.504 - 0.00344 \times \text{NEA change}$$



| NEA change $x_i$ (cal) | Fat gain $y_i$ (kg) | Predicted fat gain $\widehat{y}_i = 3.504 - 0.00344x_i$ (kg) | Residual $e_i = y_i - \widehat{y}_i$ (kg) |
|---|---|---|---|
| −94 | 4.2 | 3.504 − 0.00344 × 4.2 = 3.83 | 4.2 − 3.83 = 0.37 |
| −57 | 3.0 | 3.504 − 0.00344 × 3.0 = 3.70 | 3.0 − 3.70 = −0.70 |
| −29 | 3.7 | 3.504 − 0.00344 × 3.7 = 3.60 | 3.7 − 3.60 = 0.10 |
| 135 | 2.7 | 3.504 − 0.00344 × 2.7 = 3.04 | 2.7 − 3.04 = −0.34 |
| 143 | 3.2 | 3.504 − 0.00344 × 3.2 = 3.01 | 3.2 − 3.01 = 0.19 |
| 151 | 3.6 | 3.504 − 0.00344 × 3.6 = 2.99 | 3.6 − 2.99 = 0.61 |
| 245 | 2.4 | 3.504 − 0.00344 × 2.4 = 2.66 | 2.4 − 2.66 = −0.26 |
| 355 | 1.3 | 3.504 − 0.00344 × 1.3 = 2.28 | 1.3 − 2.28 = −0.98 |
| 392 | 3.8 | 3.504 − 0.00344 × 3.8 = 2.16 | 3.8 − 2.16 = 1.64 |
| 473 | 1.7 | 3.504 − 0.00344 × 1.7 = 1.88 | 1.7 − 1.88 = −0.18 |
| 486 | 1.6 | 3.504 − 0.00344 × 1.6 = 1.83 | 1.6 − 1.83 = −0.23 |
| 535 | 2.2 | 3.504 − 0.00344 × 2.2 = 1.66 | 2.2 − 1.66 = 0.54 |
| 571 | 1.0 | 3.504 − 0.00344 × 1.0 = 1.54 | 1.0 − 1.54 = −0.54 |
| 580 | 0.4 | 3.504 − 0.00344 × 0.4 = 1.51 | 0.4 − 1.51 = −1.11 |
| 620 | 2.3 | 3.504 − 0.00344 × 2.3 = 1.37 | 2.3 − 1.37 = 0.93 |
| 690 | 1.1 | 3.504 − 0.00344 × 1.1 = 1.13 | 1.1 − 1.13 = −0.03 |

How is the least-square regression line compared with the line $y = 3.5 - 0.004x$? The SD for these 16 least-square residuals is $\approx 0.715$kg, smaller than the SD 0.73kg of the residuals for the line $y = 3.5 - 0.004x$.



## One More Example — Men's Weight & Height

In a sample of men age 18-24, the relationship between their heights and weights is summarized as follows

average height ≈ 70", SD ≈ 3"
average weight ≈ 162 lb, SD ≈ 30 lb , $r \approx 0.5$

The scatter plot shows a linear relationship.
What is the LS regression line for predicting height from weight?

- What is $x$? What is $y$?
- slope:
- intercept:
- equation:

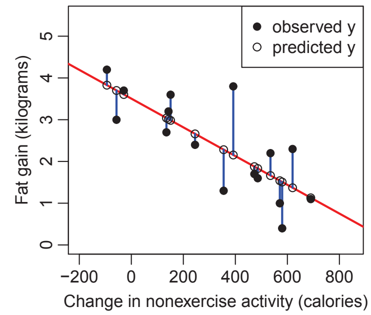
## Properties of the LS Regression Line

$$\hat{y} = \text{intercept} + \text{slope} \cdot x$$
$$= \bar{y} - \text{slope} \cdot \bar{x} + \text{slope} \cdot x$$
$$\Leftrightarrow \hat{y} - \bar{y} = \text{slope} \cdot (x - \bar{x}) = r\frac{s_y}{s_x}(x - \bar{x})$$
$$\Leftrightarrow \underbrace{\frac{\hat{y} - \bar{y}}{s_y}}_{z-\text{score of } \hat{y}} = r \cdot \underbrace{\frac{x - \bar{x}}{s_x}}_{z-\text{score of } x}$$

- The LS regression line pass through the point of the means $(\bar{x}, \bar{y})$.
  - Note the regression line may NOT pass through any of the observed data points: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Whenever $x$ increase by 1 in $z$-scores, the predicted value $\hat{y}$ only increase by $r$ in $z$-scores.
- So when $r = 0$, the predicted value $\hat{y}$ always equals the mean $\bar{y}$ regardless of the values of $x$, and the least-square regression line will be horizontal.



## Be Cautious for Extrapolation




Would you use the LS regression line

predicted fat gain = 3.504−0.00344

for predicting

- the fat gain of a young guy w/ NEA decrease 500 calories?
- the fat gain of a 70-year-old who overfed himself but w/ 0 NEA change?

A regression line can be used to make predictions for individuals. But if you have to extrapolate far from the data, or to a different group of subjects, watch out!



## Interpretation of the LS Regression Line

- The intercept is the predicted value of response for $x = 0$.
- The slope indicates how much the response changes associated with a unit change in $x$ on average (may NOT be causal).

In the young men's height and weight example, the regression line for predicting height from weight is

predicted height = 61.9" + (0.05"per lb) × (weight).

On average, a man that weighs one more pound is 0.05" taller.

- On average, a 160-pound man (age 18-24) will be 0.5" taller than a 150-pound man.
- John is 23 years old. If he puts on 10 pounds, will he become 0.5" taller?
- In this example, the intercept is meaningless since there is no man weighs 0 lb.



## There Are Two LS Regression Lines (1)

Recall the LS regression line for predicting fat gain from NEA change is

predicted fat gain = 3.504 − 0.00344 × NEA change

If a guy in the study has an NEA increase of 400 calories, his predicted fat gain is

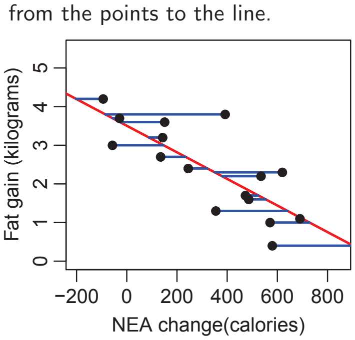
predicted fat gain = 3.504 − 0.00344 × 400 = 2.128kg

If another guy put on 2.128kg during the study, can I predict his NEA change to be 400 calories?



## There Are Two LS Regression Lines (2)

The residuals for predicting $x$ from $y$ are the horizontal distance from the points to the line.



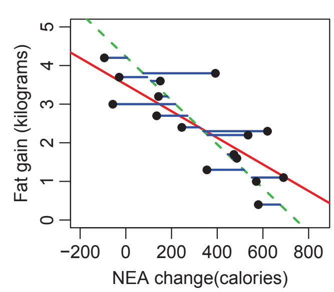

The red line is the LS regression line for predicting fat gain from NEA changes.

The red line appears to underestimate NEA changes for large fat gain, but overestimate NEA changes for low fat gain.



## There Are Two LS Regression Lines (3)

The LS regression line for predicting $x$ from $y$ is the line that minimize the sum of squared horizontal distances from the points to the line.




red solid line: predicting fat gain from NEA change
green dash line: predicting NEA change from fat gain
The two lines are different.