SKKU Biostats and Big data

Lecture 06
Linear regression

Review: Key Points

Chapter 7: Scatterplots, Correlation

- Scatterplots (direction, form, strength, outliers)
- x- and y-variables: explanatory/independent vs. response/dependent variables
- Correlation: strength and direction
- Assumptions and conditions:
  - ✓ Quantitative variables condition
  - ✓ Straight enough condition
  - ✓ No outliers condition
- Non-parametric correlations: Kendall’s tau, Spearman’s rho
- Correlation ≠ Causation
- Correlation table/matrix

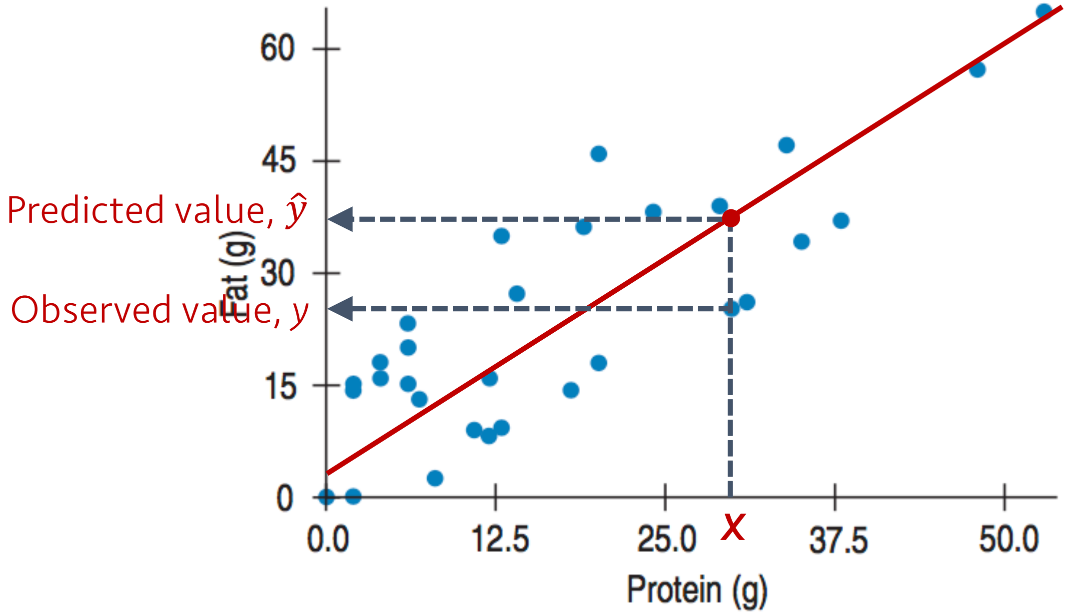
We need more than correlation!

- Correlation only tells us the strength of a linear relationship.
- Correlation doesn't tell us what the line is.
- We need a linear model*!!*

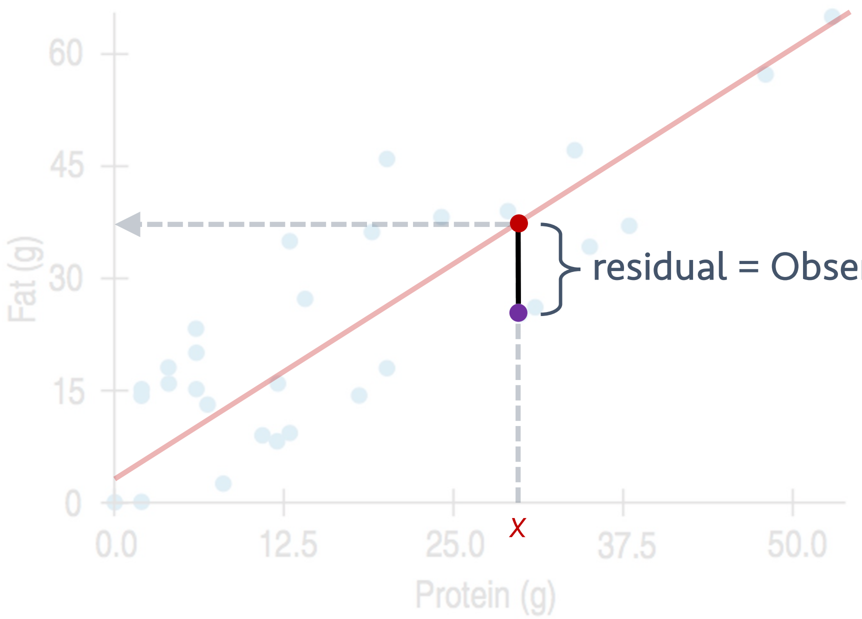
Simple regression

- Mathematical model for describing a linear relationship between an explanatory variable, x, and a response variable, y.
- It is a straight line that describes how y changes with x.
- It can be used to predict the value of y for a given value of x.

Least squares: The Line of “Best Fit”



Least squares: The Line of "Best Fit"



- Line of "best fit": the sum of the squared residuals (distance) is smallest
- $\arg\min \sum (y - \hat{y})^2 = \sum d_i^2$: *Least squares* line

c.f., deviation = Observed value (y) − mean ($\bar{y}$)

How different they are?

Linear Model

$$\hat{y} = b_0 + b_1 x.$$

- b: coefficients
- b_1: slope
- b_0: intercept

Slope, $b_1 = r\frac{s_y}{s_x}$

- Correlations do not have units, but slopes have units.
- Standard deviation as a ruler!

Intercept, $b_0 = \bar{y} - b_1\bar{x}$

Example: Sleep Study

- Sleep deprivation and study can have more errors

Errors	7	8	11	13	14	y
Hours without sleep	8	12	16	20	24	x

$$\bar{x} = 16, s_x = 6.32$$

$$\bar{y} = 10.6, s_y = 3.05$$

$$r = 0.985$$

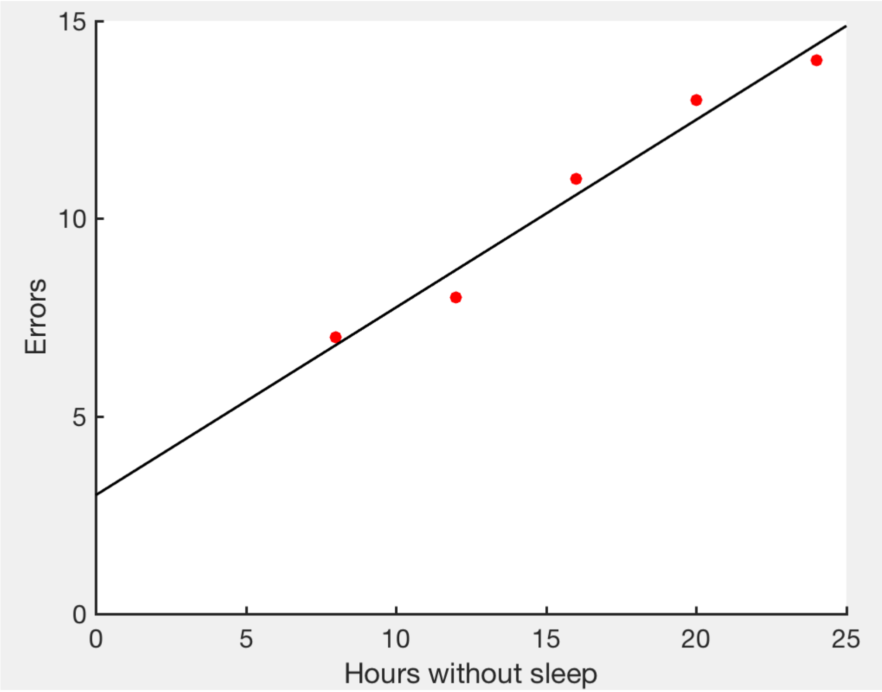
$$b_1 = r\frac{s_y}{s_x} = 0.985 \times \frac{3.05}{6.32} = 0.475$$

$$b_0 = \bar{y} - b\bar{x} = 10.6 - 0.475 \times 16 = 3$$

Least-squares regression line: $\hat{y} = b_0 + b_1 x = 3 + 0.475x$

Slide from Martin Lindquist

Example: Sleep Study



Properties

- In general the slope has units "y-units per x-units". Here errors per hour without sleep.
- The y-intercept is not always meaningful.
- The least-squares regression line always passes through the point, $(\bar{x}, \bar{y})$.
- If both the variables are standardized, the regression line is given by $\hat{z}_y = r z_x$

Least-squares regression line: $\hat{y} = b_0 + b_1 x = 3 + 0.475x$

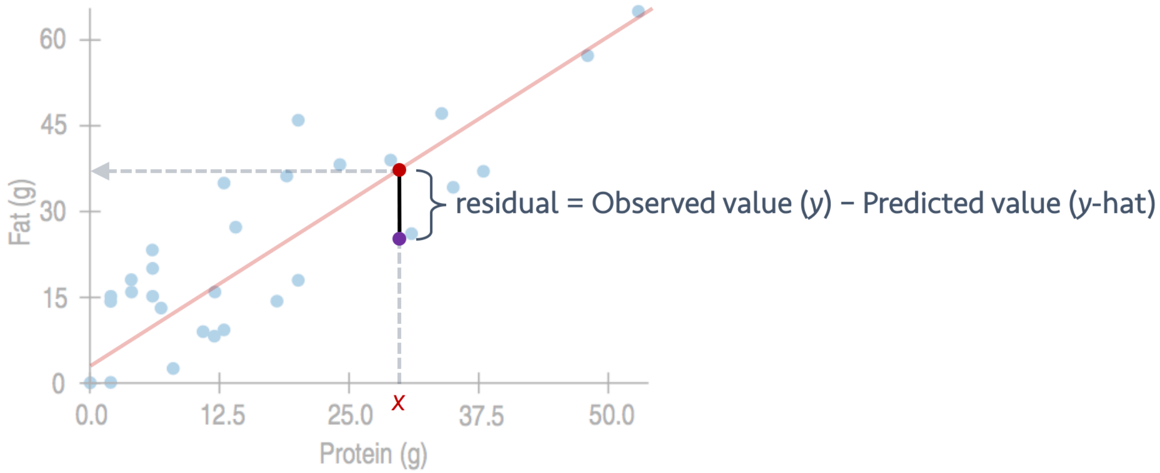
Slide from Martin Lindquist

Quiz 06-1

https://forms.gle/iUgs16PnQYA8iFGP7

Examining the residuals

- Residuals are defined as:
- $e = y - \hat{y}$



Examining the residuals

- Residuals are defined as:
- $e = y - \hat{y}$

- In least square regression, the sum of the residuals is always zero.
- The residuals are the variation in the data that has not been modeled.
  - DATA = MODEL + RESIDUAL

$$\hat{y} = b_0 + b_1 x$$

$$y = b_0 + b_1 x + e$$

- A residual plot is a scatter plot of the residuals against x or $\hat{y}$.
- When studying the residual plot we hope to see NO pattern.

Slide partly from Martin Lindquist

Examining the residuals: Sleep study

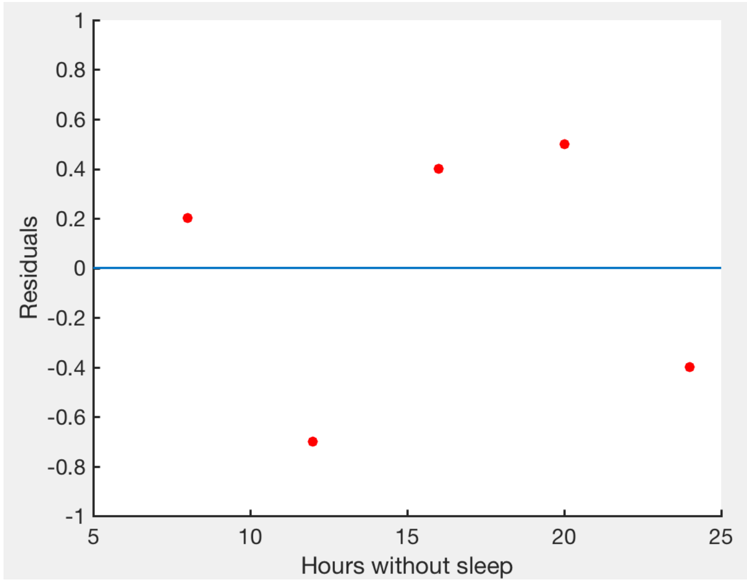
Errors	7	8	11	13	14
Hours without sleep	8	12	16	20	24

```
>> residuals = y-(3+0.475*x);
>> sum(residuals)

ans =

   2.6645e-15

>> scatter(x, residuals)
```



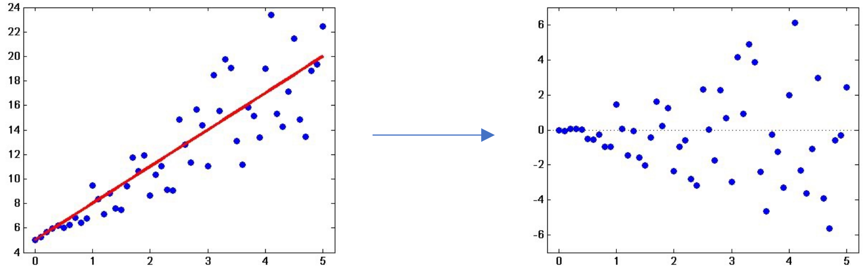
- The sum of the residuals is zero.
- should be the most boring scatterplot you've ever seen!
- shouldn't have any interesting features, direction or shape
- should stretch horizontally, with about same amount of scatter throughout
- No bends, no outliers

Residual standard deviation

- s_e
- tells us how much the points spread around the regression line.

$$s_e = \sqrt{\frac{\sum e^2}{n-2}}$$

- Revisit: Correlation assumptions and conditions
  - ✓ Quantitative variables condition
  - ✓ Straight enough condition
  - ✓ No outliers condition
- In regression, one more condition:
  - ✓ **Does the Plot Thicken? Condition**
    - Equal variance assumption
    - The spread around the line should not increase as x or the predicted values increase.



Regression Assumptions and Conditions

- Quantitative Variable Condition
- Straight Enough Condition
- Outlier Condition
- Does the Plot Thicken? Condition

Examining residual plots:

- No bends (Straight Enough Condition)
- No outlier (Outlier Condition): “examine points with large residuals”
- No changes in the spread (Does the Plot Thicken? Condition)

Quiz 06-2

https://forms.gle/5ZnesK8RNAtCBfp66

Assessing regression model: R^2

- Correlation: strength and direction
- To evaluate how well a regression model does, direction won't matter that much.
- R^2: ranges between 0 and 1
- tells us the fraction of the data's variation accounted for by the model

- $R^2 = 1 - \frac{Sum\ of\ squared\ residuals}{Sum\ of\ squared\ deviation\ from\ the\ mean} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$
  - In the linear model, R^2 is same with r^2.
  - $1-r^2$: the fraction of the original variation left in the residuals

- How big should R^2 be? *It depends! Data type, field, etc.*
- What is more important between b and R^2? *It depends! Data type, field, research question, etc.*

Quiz 06-3

https://forms.gle/ApD8G6CScieuYXQy7

Predicting in the Other Direction

- Predicting y with x and predicting x with y are different!
- What we're minimizing when predicting y with x? $\sum(y-\hat{y})^2 = \sum(y-(b_0+b_1x))^2$
- What we need to minimize when predicting x with y, then?
  - $\sum(x-\hat{x})^2 = \sum(x-(b'_0+b'_1y))^2$
  - where $b'_1 = r\frac{s_x}{s_y}$, compared to $b_1 = r\frac{s_y}{s_x}$
- What if we're using standardized values (z-scores) in regression?
  - They are same!

Quiz 06-4

https://forms.gle/3Bofp5aR3PUzJHXz8

Key Points

Chapter 8: Linear Regression

- residual = Observed value (y) – Predicted value ($\hat{y}$)
- Line of "best fit": $\arg\min \sum (y - \hat{y})^2 = \sum d_i^2$: *Least squares* line
- $\hat{y} = b_0 + b_1 x$. Slope, $b_1 = r \frac{s_y}{s_x}$ Intercept, $b_0 = \bar{y} - b_1 \bar{x}$
- Residuals $e = y - \hat{y}$
- DATA = MODEL + RESIDUAL: $y = b_0 + b_1 x + e$
- Residual plot should show no interesting pattern.
- $R^2 = 1 - \frac{Sum\ of\ squared\ residuals}{Sum\ of\ squared\ deviation\ from\ the\ mean} = 1 - \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2}$. In the linear model, R^2 is same with r^2.
- Predicting y with x and predicting x with y are different!