Lecture 22
Inferences for Regression

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

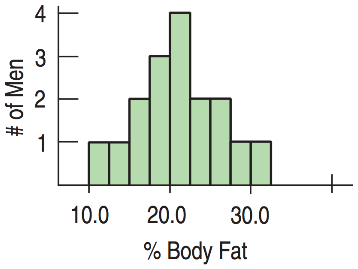
Review: Key Points

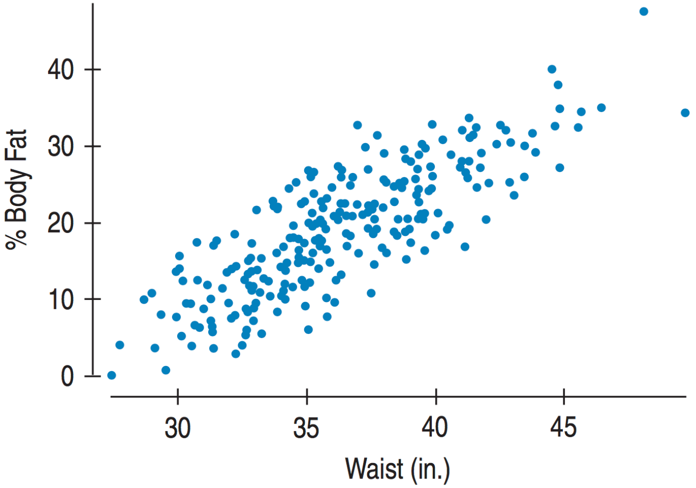
Chapter 26: Comparing Counts

- **Goodness-of-fit tests:** $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$
- Assumption and conditions:
  - counted data condition, independence assumption, expected cell frequency condition
- **Chi-square distribution:** only positive, right skewed, mode: *df*-2, mean: *df*
- Chi-square test for a one-way count table
- Chi-square test for a two-way table: Chi-square test of homogeneity
- Chi-square test for a contingency table: Chi-square test of independence

Example: predicting %body fat

- Measuring body fat is tedious and expensive.
- Can we predict %Body fat from easily measurable variables, such as *Height*, *Weight*, or *Waist* size?



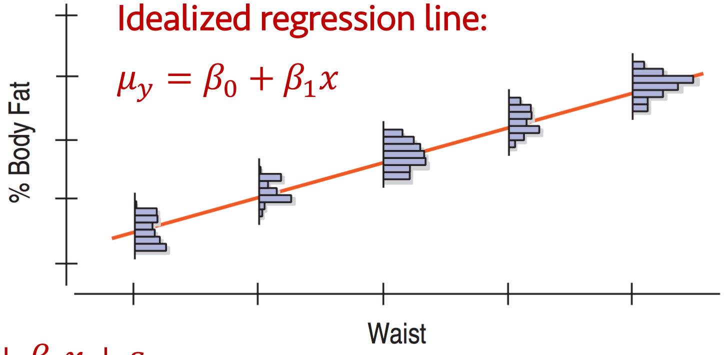


$$\widehat{\%Body\ Fat} = -42.7 + 1.7\, Waist.$$

- $\hat{y} = b_0 + b_1 x$
- How useful is this model?
- What can this model tell us beyond the 250 men in this study?

Idealized regression line:

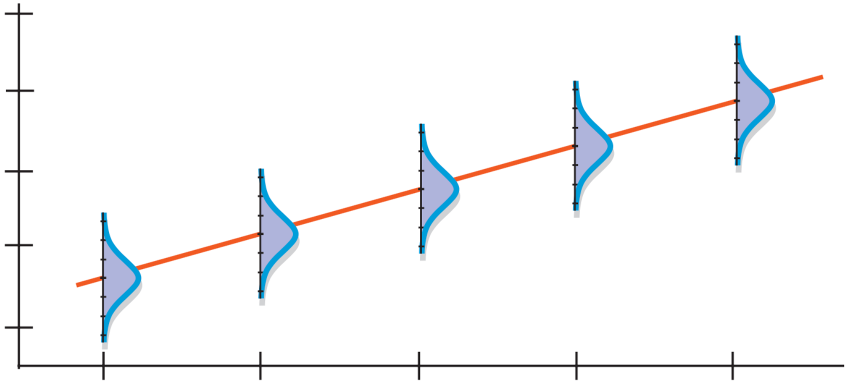
$\mu_y = \beta_0 + \beta_1 x$



$y = \beta_0 + \beta_1 x + \varepsilon$

Assumptions and Conditions

- **Linearity Assumption**
  - Quantitative data condition
  - Straight enough condition
- **Equal Variance Assumption**
  - Does the Plot Thicken? Condition
- **Normal Population Assumption**
  - Nearly Normal condition
  - Outlier condition
- **Independence Assumption**
  - Errors should be independent of each other
  - Examine residuals for evidence of patterns, trends, or clumping, etc.
  - If the data are related to time (e.g., time-series), plot residuals against time.



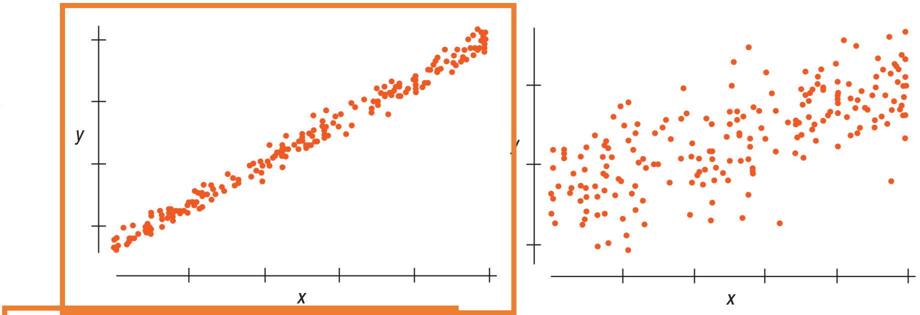
The order of examining assumptions and conditions

1. Make a scatterplot of the data to check the Straight Enough Condition.
   - If the relationship is curved, try re-expressing the data. Or stop.
2. If the data are straight enough, fit a regression and find the residuals, e, and predicted values, $\hat{y}$
3. Make a scatterplot of the residuals against *x* or against the predicted values. This plot should have **no pattern.**
   - Check for any bend (which would suggest that the data weren't all that straight after all),
   - for any thickening (or thinning)
   - for any outliers. (If there are outliers, and you can correct them or justify removing them, do so and go back to step 1, or consider performing two regressions—one with and one without the outliers.)
4. If the data are measured over time, plot the residuals against time to check for evidence of patterns that might suggest they are not **independent**.
5. If the scatterplots look OK, then make a histogram and Normal probability plot of the residuals to check the Nearly Normal Condition.
6. If all the conditions seem to be reasonably satisfied, go ahead with inference.

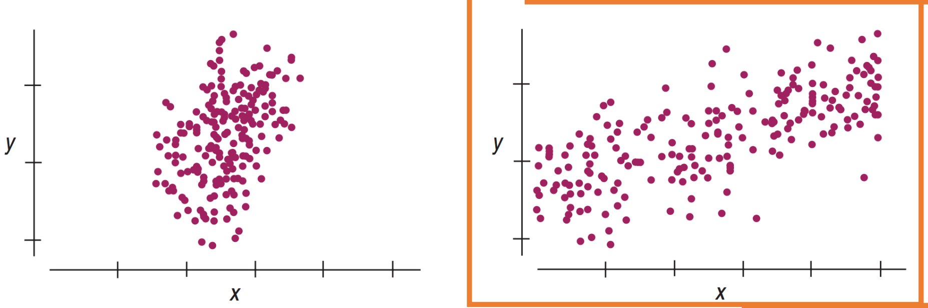
Intuition about regression inference

The factors that affect how much the slope (and intercept) vary from sample to sample (stability of beta)

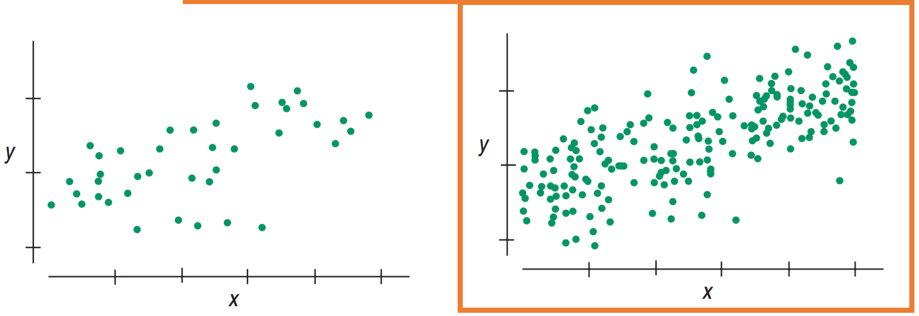
- Spread around the line
  - Residual standard deviation, $s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$



- Spread of the *x*'s



- Sample size



Standard Error for the Slope

- Spread around the line: s_e
- Spread of the x's: s_x
- Sample size: n

- Standard error for the regression slope:
  - $SE(b_1) = \frac{s_e}{\sqrt{n-1}s_x}$, where $s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$, $s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
  - Similar to inferences about means, $\frac{b_1-\beta_1}{SE(b_1)} \sim t_{n-2}$
  - Same for Intercept:
    - $\frac{b_0-\beta_0}{SE(b_0)} \sim t_{n-2}$

Regression inference

- Usual null hypothesis: Slope = 0

- H$_0$: $\beta_1 = 0$,
  - $t_{n-2} = \frac{b_1 - 0}{SE(b_1)}$

- 95% confidence interval for β
  - $b_1 \pm t^*_{n-2} \times SE(b_1)$

Standard Errors for Predicted Values

- Do we want to know the mean *%Body Fat* for *all* men with a *Waist size* of 38 inches?
- Do we want to estimate the *%Body Fat* for a particular man with a *Waist size* of 38 inches?

- $\hat{y}_\nu \pm t^*_{n-2} \times SE$
  - But different *SE*s for different questions
    - Mean: $SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$

Standard Errors for Predicted Values

- Do we want to know the mean *%Body Fat* for *all* men with a *Waist size* of 38 inches?
- Do we want to estimate the *%Body Fat* for a particular man with a *Waist size* of 38 inches?

- $\hat{y}_\nu \pm t^*_{n-2} \times SE$
  - But different *SE*s for different questions
    - Mean: $SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$

Less certain of the slope, less certain of our predictions

Farther the prediction for x_ν from $\bar{x}$, lower certainty

More spread around the line, lower certainty

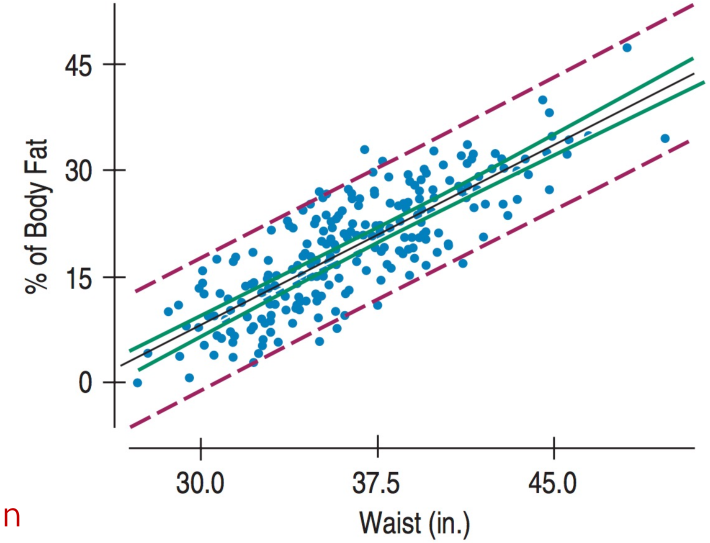
More data, higher certainty

Standard Errors for Predicted Values

- Do we want to know the mean *%Body Fat* for *all* men with a *Waist size* of 38 inches?
- Do we want to estimate the *%Body Fat* for a particular man with a *Waist size* of 38 inches?

- $\hat{y}_\nu \pm t^*_{n-2} \times SE$
  - But different *SE*s for different questions
    - Mean: $SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$
    - Individual: $SE(\hat{y}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$

The variation of individuals around the predicted mean



Key Points

Chapter 27: Inferences for Regression

- $\mu_y = \beta_0 + \beta_1 x, y = \beta_0 + \beta_1 x + \varepsilon$
- Assumption and conditions:
  - Linearity Assumption, Equal Variance Assumption, Normal Population Assumption, Independence Assumption
- $SE(b_1) = \frac{s_e}{\sqrt{n-1}s_x}$, where $s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$, $s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
- Hypothesis test: H_0: $\beta_1 = 0$, $t_{n-2} = \frac{b_1 - 0}{SE(b_1)}$
- 95% confidence interval for β: $b_1 \pm t_{n-2}^{*} \times SE(b_1)$
- Standard errors for predicted values: $\hat{y}_\nu \pm t_{n-2}^{*} \times SE$
  - Mean: $SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$, Individual: $SE(\hat{y}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$