Announcements

- Written homework 4 will be posted today!
  - In general, make sure you submit the correct assignments to the correct stops in Gradescope!
- Lab 4 write-up due tonight!
- I'm adding a Multiple Regression Lab on Wednesday, so bring your laptops
- Test 1 a week from this Friday
  - I'll aim to get study materials up by the weekend
  - If you haven't done them, you do have the practice problems from each chapter
- Read Ch 2.1 and 2.2 for Friday

Warm Up

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

	Estimate	Std. Error	t value	Pr($>|t|$)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00

$R^2_{adj} = 0.057$

A) higher

B) lower

C) the same

D) it is impossible to tell from this information

Warm Up

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00

$R^2_{adj} = 0.057$

A) higher

B) lower

C) the same

D) it is impossible to tell from this information

Model Assumptions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

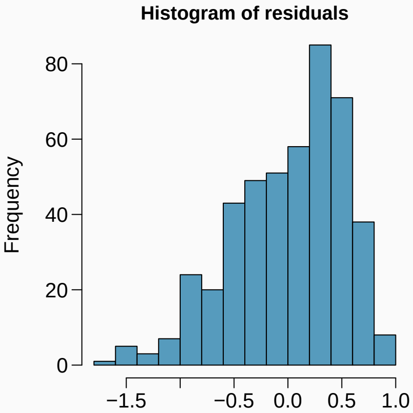
The least squares model depends on the following conditions:

- □ Residuals are nearly normal (unimodal and symmetric)
- □ Residuals have constant variability
- □ Residuals are independent
- □ Each variable is linearly related to the response

We generally use graphical methods to check these.

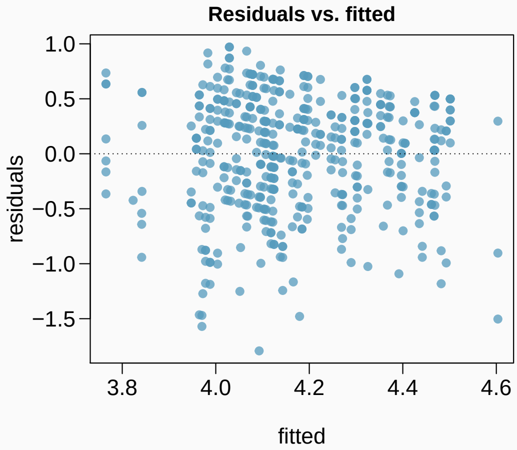
Condition 1: Nearly normal residuals

- Use histogram of residuals
- Mostly concerned with being unimodal and no odd outliers

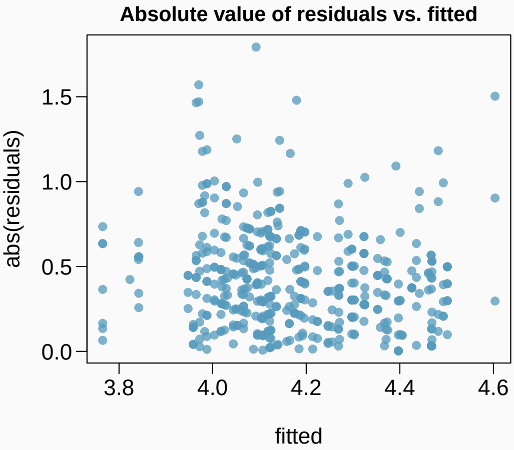


Condition 2: Constant Residual Variability

Use a scatterplot of residuals and/or absolute value of residuals vs predicted values.



Residuals vs. fitted



Absolute value of residuals vs. fitted

Condition 2: Why vs Predicted?

- When we did simple linear regression, we checked the constant variance using a plot of ***residuals vs x***.
- With multiple regression, we check constant variance using a plot of ***residuals vs predicted***.

Why the difference?

Condition 2: Why vs Predicted?

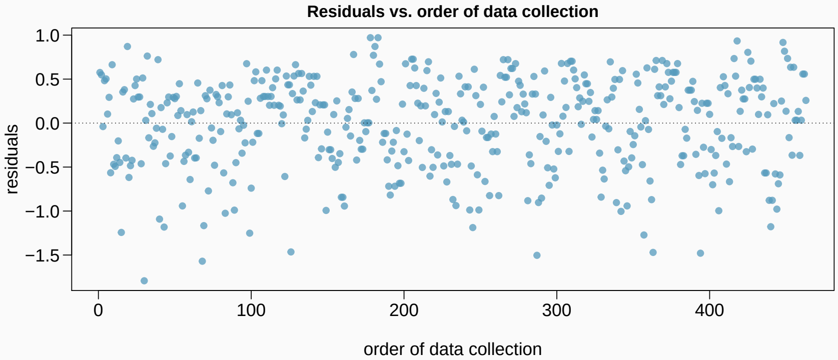
- When we did simple linear regression, we checked the constant variance using a plot of ***residuals vs x***.
- With multiple regression, we check constant variance using a plot of ***residuals vs predicted***.

Why the difference?

- In multiple regression there are many explanatory variables, so a plot of residuals vs one of them wouldn't give a full picture.

Condition 3: Independent Residuals

Scatterplot of residuals vs order of data collection.



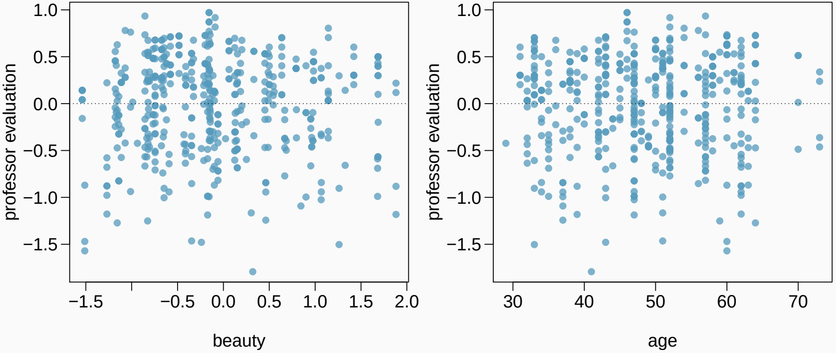
Helps check against related effects or time-of-day effects.

Condition 3: More on Independent Residuals

- Checking for independent residuals allows us to indirectly check for independent observations
- If observations and residuals are independent, we would not expect to see any trend in the residuals vs order of data scatterplot
- This condition is often violated with time-series data. Such data require more advanced time series regression techniques

Condition 4: Linear Relationships

Scatterplot of residuals vs ***each*** numerical explanatory variable



Upon Failure

If your multiple regression model fails one of the diagnostics:

- See if you can fix it!
  - Did you add an extra explanatory variable that it turns out was non-linear?
  - Can you try a different model that might better match your data?
- In not, report the model of note its shortcomings
  - Failings can indicate important parts of the data as well!

Conditional Probability

Relapse Example

Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard cocaine treatment), and placebo. Results of the study are summarized below.

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Source: `http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm`

Marginal Probability

What is the probability that a patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Marginal Probability

What is the probability that a patient relapsed?

		no	
	relapse	relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	*48*	24	*72*

$$\text{P(relapsed)} = \frac{48}{72} \approx 0.67$$

Joint Probability

What is the probability that the patient received the antidepressant (desipramine) and relapsed?

		no	
	relapse	relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Joint Probability

What is the probability that the patient received the antidepressant (desipramine) and relapsed?

		no	
	relapse	relapse	total
desipramine	*10*	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	*72*

P(relapsed and desipramine) $= \frac{10}{72} \approx 0.14$

Conditional Probability

Conditional Probability

The conditional probability of the outcome of interest A given condition B is calculated as

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(\text{relapse}|\text{desi}) = \frac{P(\text{relapse and desi})}{P(\text{desi})}$$

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Conditional Probability

Conditional Probability

The conditional probability of the outcome of interest A given condition B is calculated as

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$\begin{aligned} P(\text{relapse}|\text{desi}) &= \frac{P(\text{relapse and desi})}{P(\text{desi})} \\ &= \frac{10/72}{24/72} \\ &= 0.42 = \frac{10}{24} \end{aligned}$$

	relapse	no relapse	total
desipramine	*10*	14	*24*
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Understanding Check

What is the probability that the patient received lithium given that the patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

A) 0.25

B) 0.375

C) 0.75

D) 0.9

Understanding Check

What is the probability that the patient received lithium given that the patient relapsed?

		no	
	relapse	relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

A) 0.25

B) 0.375

C) 0.75

D) 0.9

Conditional Probabilities and Independence

Generally, if $P(A|B) = P(A)$, then the events A and B are said to be independent.

- Conceptually: Giving B doesn't tell us anything about A.
- Mathematically: We know that if A and B are independent, then $P(A \text{ and } B) = P(A) \times P(B)$. Thus:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

Breast Cancer Screening

- American Cancer Society estimates that about 1.7% of women have breast cancer.

  Source: `http://www.cancer.org/cancer/cancerbasics/cancer-prevalence`

- Susan G. Komen For The Cure Foundation states that mammography correctly identifies about 78% of women who truly have breast cancer.

  Source: `http://ww5.komen.org/BreastCancer/AccuracyofMammograms.html`

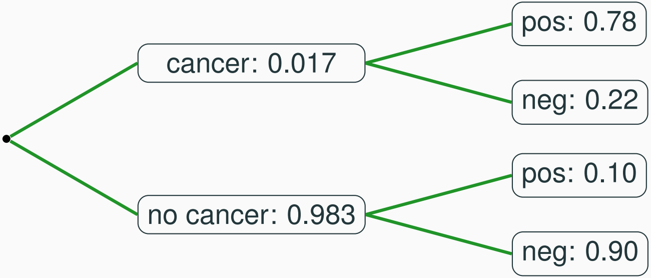
- An article published in 2003 suggests that up to 10% of all mammograms result in false positives for patients who do not have cancer.

  Source: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940`

Note: These percentages are approximate and very difficult to estimate.

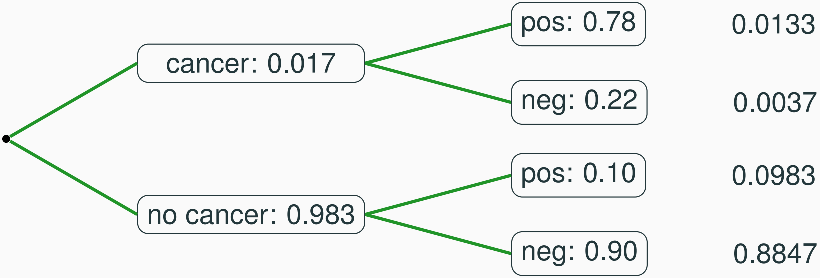
Inverting Probabilities

If a mammogram yields a positive result, what is the probability that a patient actually has cancer?



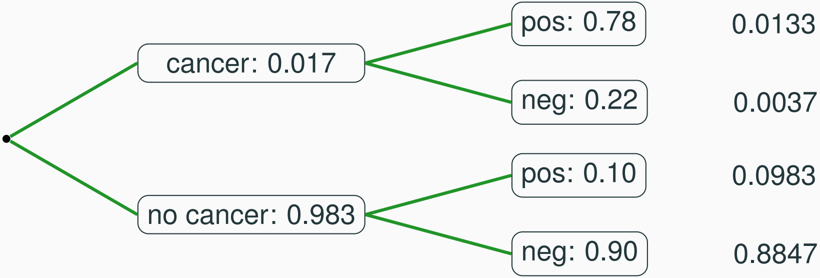
Inverting Probabilities

If a mammogram yields a positive result, what is the probability that a patient actually has cancer?



Inverting Probabilities

If a mammogram yields a positive result, what is the probability that a patient actually has cancer?



cancer: 0.017

pos: 0.78 0.0133

neg: 0.22 0.0037

no cancer: 0.983

pos: 0.10 0.0983

neg: 0.90 0.8847

$$P(C|+) = \frac{P(C \text{ and } +)}{P(+)} = \frac{0.0133}{0.0133 + 0.0983} = 0.12$$