Review for Final Exam
Stat 205: Statistics for the Life Sciences

Tim Hanson, Ph.D.

University of South Carolina

Logistics...

* Open book, *not open notes*. Bring a calculator.
* You can put post-it notes in your book.
* Thursday, April 28, 9am–noon. *Be on time*.
* 3 regular problems, each worth 20 points: 60 total.
* Rest of today's lecture is reviewing this material.
* 1 problem is short answer spanning entire course.
* JeanMarie's office hours next week: Monday 9-11, Tuesday/Thursday 11-12, & Wednesday 11-1 in LeConte 215A. Dr. Hanson's office hours: Monday/Wednesday 1:30-2:30, & Tuesday 1-2.

12.1, 12.2, 12.3, 12.4: Linear regression

* Have scatterplot of n paired values (x_1, y_1), (x_2, y_2),...,(x_n, y_n).
* Theoretical model:

$$\underbrace{Y_i}_{\text{observed}} = \underbrace{\beta_0 + \beta_1 x_i}_{\text{unknown trend}} + \underbrace{\epsilon_i}_{\text{slop}}$$

* We use data (x_1, y_1), (x_2, y_2),...,(x_n, y_n) to obtain the *fitted line*

$$Y = b_0 + b_1 x,$$

where b_0 and b_1 are the *least squares estimates* of β_0 and β_1:

$$\begin{aligned} b_1 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ b_0 &= \bar{y} - b_1 \bar{x} \end{aligned}$$

Linear regression

* Estimate of var(ϵ_i) is

$$s_{y|x} = \sqrt{\frac{SSE(resid)}{n-2}},$$

where

$$SS(resid) = \sum(y_i - b_0 - b_1 x_i)^2 = \sum(y_i - \hat{y}_i)^2.$$

* $s_y = \sqrt{\frac{1}{n-1}\sum(y_i - \bar{y})^2}$ is overall variability of $y_1, \ldots, y_n$ around the mean $\bar{y}$. $s_{y|x} = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}$ is overall variability of $y_1, \ldots, y_n$ around the line $b_0 + b_1 x$.
* $s_{y|x} < s_y$. How *much* smaller tells you how "well" the line is working to explain the data.
* HW: 12.3, 12.7, 12.9.

Linear regression

* If we assume slop $\epsilon_1, \ldots, \epsilon_n$ are *normal* then

$$SE_{b_1} = \frac{s_{y|x}}{\sqrt{\sum (x_i - \bar{x})^2}}.$$

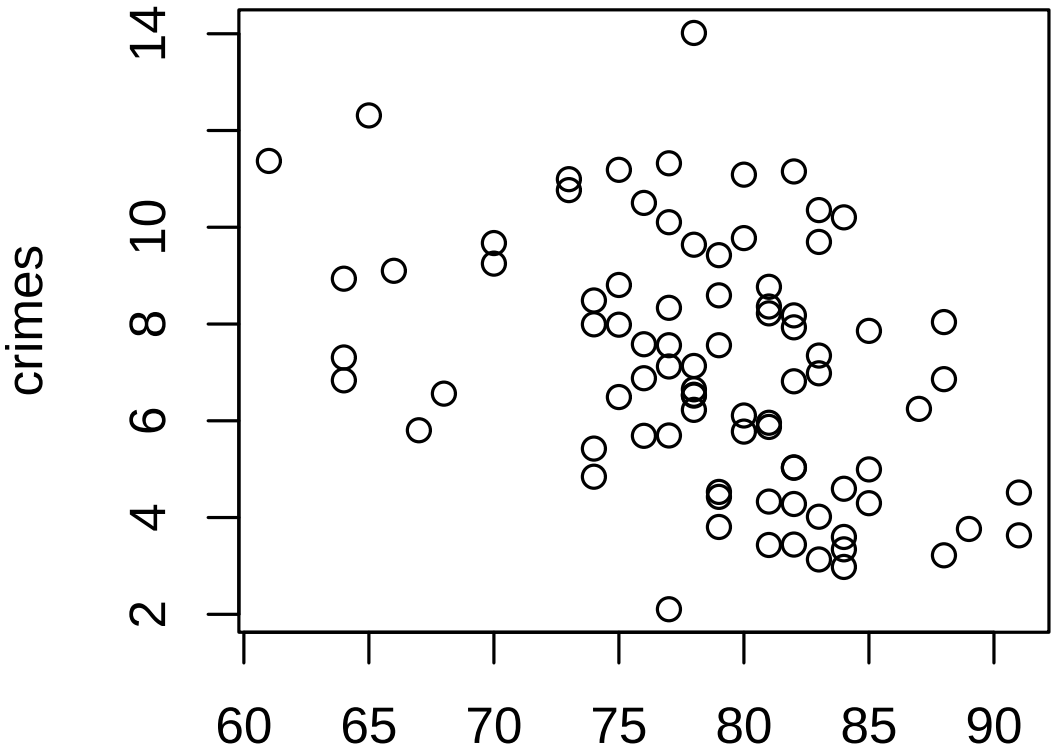
* Inference for β_1:
  (a) $(1-\alpha)100\%$ confidence interval for β_1 has endpoints

$$b_1 \pm t_{1-\alpha/2} SE_{b_1},$$

  where $df = n - 2$.
  (b) To test $H_0 : \beta_1 = 0$ at level α, see if confidence interval from (a) includes zero. If not, reject $H_0 : \beta_1 = 0$.
* If you reject $H_0 : \beta_1 = 0$ then x and y are *significantly, linearly related*.
* 5 statistics needed: $\bar{x}$, $\bar{y}$, $\sum (x_i - \bar{x})^2$, $\sum (x_i - \bar{x})(y_i - \bar{y})$, $SS(resid) = \sum (y_i - b_0 - b_1 x_i)^2$.
* HW: 12.16, 12.17, 12.19, 12.23(a).

Example

A criminologist studying the relationship between level of education and crime rate in medium-sized U.S. counties collected data from a random sample of $n = 84$ counties. Y is the crime rate (crimes per 100 people) and X is the percentage of individuals in the county having at least a high-school diploma. Here's a scatterplot:



Example, continued...

For these data, $n = 84$, $\bar{y} = 7.111$, $\bar{x} = 78.60$,
$\sum(x_i - \bar{x})(y_i - \bar{y}) = -547.9$, $\sum(x_i - \bar{x})^2 = 3212$, $SS(resid) = 455.3$.

* The least squares estimates b_0 and b_1 for regressing crime rate on percentage of high-school graduates are computed

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{-547.9}{3212} = -0.171.$$

$$b_0 = \bar{y} - b_1\bar{x} = 7.111 - (-0.171)78.60 = 20.5.$$

* For every percent increase in those receiving a high-school diploma, the crime rate drops 0.17 per 100 people, on average.
* For Richland County, $X = 85.2\%$. We predict Richland's crime rate to be

$$20.5 - 0.171(85.2) = 5.9 \text{ people / } 100.$$

Example, continued...

* A 95% confidence interval for β_1 is computed

$$SE_{b_1} = \frac{s_{y|x}}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{\sqrt{SS(resid)/(n-2)}}{\sqrt{\sum(x_i - \bar{x})^2}}$$
$$= \frac{\sqrt{455.3/(84-2)}}{\sqrt{3212}} = 0.0416.$$

$$b_1 \pm t_{0.975} SE_{b_1} = -0.171 \pm 1.99 \times 0.0416 = (-0.25, -0.088).$$

Note that $df = 84 - 2 = 82$ for the t distribution; I rounded down to $df = 80$ to use the table in the back of the book.

* Question Do we reject $H_0 : \beta_1 = 0$ at the 5% significance level? Why or why not? What does this tell you about the relationship between crime and education? Answer We reject $H_0 : \beta_1 = 0$ at the 5% significance level because a 95% confidence interval for β_1 does not include zero. There is a significant, negative association between crime rate and graduating high school.

10.7 & 10.9: 2×2 tables

* Want to compare proportions/probabilities across two groups, e.g. proportion of diabetics for obese versus normal weights.
* Have two independent binomial samples from two populations $y_1 \sim \text{binom}(n_1, p_1)$ & $y_1 \sim \text{binom}(n_2, p_2)$:

	Group 1	Group 2
With attribute	y_1	y_2
Without attribute	$n_1 - y_1$	$n_2 - y_2$
total	n_1	n_2

* Estimate p_1 and p_2 by $\hat{p}_1 = y_1/n_1$ and $\hat{p}_2 = y_2/n_2$.
* The estimated difference in proportions is $\hat{p}_1 - \hat{p}_2$.
* The estimated relative risk is $\hat{p}_1/\hat{p}_2$.
* The estimated odds ratio is $\hat{\theta} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{y_1 \times (n_2 - y_2)}{y_2 \times (n_1 - y_1)}$.

2 × 2 tables

* 95% confidence interval for $p_1 - p_2$. Let

$$\tilde{p}_1 = \frac{y_1 + 1}{n_1 + 2} \text{ and } \tilde{p}_2 = \frac{y_2 + 1}{n_2 + 2},$$

and

$$SE_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}.$$

Then a 95% CI for $p_1 - p_2$ is given by

$$\tilde{p}_1 - \tilde{p}_2 \pm 1.96 \; SE_{\tilde{p}_1 - \tilde{p}_2}.$$

* To test $H_0 : p_1 = p_2$ at 5% significance level, see if confidence interval above includes zero. If not, then *reject*.

2 × 2 tables

* 95% confidence interval for $\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$.
* First get 95% CI for $\log(\theta)$:

$$\log \hat{\theta} = \log \left(\frac{y_1(n_2 - y_2)}{y_2(n_1 - y_1)} \right).$$

The standard error of the log-odds ratio is

$$SE_{\log(\hat{\theta})} = \sqrt{\frac{1}{y_1} + \frac{1}{y_2} + \frac{1}{n_1 - y_1} + \frac{1}{n_2 - y_2}}.$$

The 95% CI for the log-odds ratio from a 2×2 table is

$$\log \hat{\theta} \pm 1.96 \; SE_{\log(\hat{\theta})}.$$

First, compute this interval, then exponentiate both numbers to get a confidence interval for θ.

* To test $H_0 : \theta = 1$ at 5% significance level, see if confidence interval above includes **one**. If not, then *reject*.

2×2 tables

* We accept/reject $H_0 : p_1 = p_2$ and $H_0 : \theta = 1$ at the same time; one implies the other. If we reject, then there is a significant *association* between the grouping variable and the probability of the event.
* The interpretation for the odds ratio can "flip." Very, very important. Relative risks do not have this property. See online notes and pp. 446–450.
* For *rare events*, the odds ratio equals the relative risk.
* HW: 10.57, 10.58, 10.59 (for all problems also compute relative risk, odds ratio, and 95% CI for the OR; formally test $H_0 : OR = 1$), 10.68, 10.69, 10.70.

Example, problem 10.59

The data are

	Bed rest	control
Preterm delivery	32	20
Normal delivery	73	87
total	105	107

Comparing proportions: Compute $\hat{p}_1 = 32/105 = 0.305$ and $\hat{p}_2 = 20/107 = 0.187$.

* We estimate the difference $p_1 - p_2$ to be $\hat{p}_1 - \hat{p}_2 = 0.305 - 0.187 = 0.118$. The probability of preterm *increases* by 0.12 for those on complete bed rest.
* We estimate the relative risk to be $\hat{p}_1/\hat{p}_2 = 0.305/0.187 = 1.63$. The probability of preterm increases by 63% for bed rest.
* We estimate the odds ratio to be $\hat{\theta} = y_1(n_2 - y_1)/[y_2(n_1 - y_1)] = [32 \times 87]/[20 \times 73] = 1.91$. The odds of preterm almost doubles under complete bed rest.

Example, problem 10.59

95% confidence interval for $p_1 - p_2$:

Let

$$\tilde{p}_1 = \frac{32+1}{105+2} = 0.308 \text{ and } \tilde{p}_2 = \frac{20+1}{107+2} = 0.192,$$

$$SE_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{\frac{0.308(1-0.308)}{105+2} + \frac{0.192(1-0.192)}{107+2}} = 0.058.$$

Then a 95% CI for $p_1 - p_2$ is given by

$$0.308 - 0.192 \pm 1.96\ (0.058) = (0.001, 0.230).$$

We are 95% confident that complete bed rest increases the probability of preterm delivery from 0.1% to 23%.

Example, problem 10.59

95% confidence interval for θ:

$$se_{\log \hat{\theta}} = \sqrt{\frac{1}{32} + \frac{1}{20} + \frac{1}{73} + \frac{1}{87}} = 0.326.$$

A 95% CI for $\log \theta$ is

$$0.645 \pm 1.96(0.326) = (0.006, 1.285).$$

Exponentiating both sides we get $(1.006, 3.614)$. We are 95% confident that complete bed rest increases the odds of preterm delivery by 1% to 260%. Since the 95% CI for θ (barely) does not include one, we reject $H_0 : \theta = 1$ at the 5% level: there is a signficant statistical association between bed rest and preterm delivery.

12.7 (pp. 582–585) Logistic regression

* Have n paired values (x_1, y_1), (x_2, y_2),...,(x_n, y_n).
* Theoretical model:

$$\Pr\{Y_i = 1\} = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

* A statistical package uses data (x_1, y_1), (x_2, y_2),...,(x_n, y_n) to obtain (b_0, b_1), giving the *fitted probability function*

$$\hat{p}(x) = \frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)}$$

where b_0 and b_1 are estimates of β_0 and β_1:

* These estimates are given to you from a computer package such as SAS, Minitab, or web-based applets. Here's an Excel version: http://udel.edu/~mcdonald/statlogistic.html

* Table looks like

Parameter	Est.	S.E.	Test stat.	p-value
Intercept	b_0	SE_{b_0}	$\frac{b_0}{SE_{b_0}}$	Tests $H_0 : \beta_0 = 0$
continuous predictor (slope)	b_1	SE_{b_1}	$\frac{b_1}{SE_{b_1}}$	Tests $H_0 : \beta_1 = 0$

* Testing $H_0 : \beta_1 = 0$ tests whether there is *an association* between the predictor x and the response Y. If we reject $H_0 : \beta_1 = 0$ then *there is a significant association* between the two.
* e^{b_1} is an estimate of how the odds of $Y = 1$ change when x is increased by one unit. A 95% confidence interval for this *odds ratio* is $(e^{b_1 - 1.96 SE_{b_1}}, e^{b_1 + 1.96 SE_{b_1}})$.

Example

80 students were polled in Stat 205 Spring 2011 on whether they lived on campus, and their age (years). A logistic regression model was fit in Minitab yielding:

```
Variable   Value  Count
residence  on        36  (Event)
           off       44
           Total     80

Logistic Regression Table

Predictor       Coef    SE Coef       Z      P
Constant     18.7198    5.24220    3.57  0.000
age        -0.961319   0.269191   -3.57  0.000
```

First, $b_1 < 0$ and we reject $H_0 : \beta_1 = 0$ at the 5% level because $0.000 < 0.05$: *there is a significant, negative association between living on campus and age*.

Question what if the p-value was 0.17?

Example

```
Variable   Value  Count
residence  on        36  (Event)
           off       44
           Total     80

Logistic Regression Table

Predictor       Coef    SE Coef       Z      P
Constant     18.7198    5.24220    3.57  0.000
age         -0.961319  0.269191   -3.57  0.000
```

Next, for every year increase in age, we estimate that the odds of living on campus changes by $e^{-0.961} \approx 0.38$. *For every year increase in age, the odds of living on campus decreases by more than half*. We can compute a 95% confidence interval as

$$\exp(-0.961 - 1.96(0.269)) \approx 0.23 \text{ and } \exp(-0.961 + 1.96(0.269)) \approx 0.65.$$

With 95% confidence, the odds of living on campus change by a factor ranging from 0.23 to 0.65 for every year increase in age.

The probability of living off campus as a function of age is given by

$$\hat{p}(x) = \frac{\exp(18.72 - 0.96\ x)}{1 + \exp(18.72 - 0.96\ x)}.$$

This function and the raw proportions at ages 18, 19, 20, 21, 22, 23, 24, 25, 33, 37, 43 are plotted below: