Chapter 12: Linear regression II

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

12.4 The regression model

- We assume the underlying model with Greek letters (as usual)

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

- For each subject i we see x_i and $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- β_0 is the population intercept.
- β_1 is the population slope.
- ϵ_i is the ith error, we assume these are $N(0, \sigma_e)$.
- We don't know any of β_0, β_1, or σ_e.

Visualizing the model

- $\mu_{y|x} = \beta_0 + \beta_1 x$ is mean response for everyone with covariate x.
- σ_e is constant variance. Variance doesn't change with x.
- Example 12.4.4, pretend *we know* that the mean weight $\mu_{y|x}$ given height x is

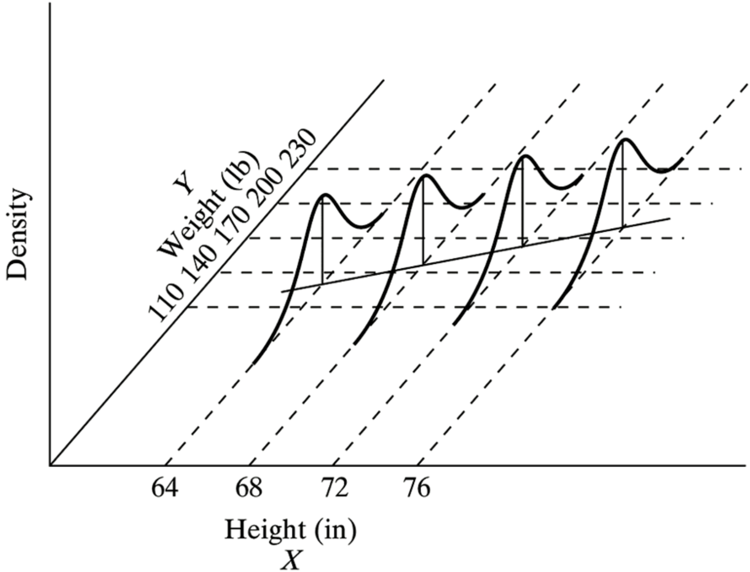
$$\mu_{y|x} = -145 + 4.25x \text{ and } \sigma_e = 20.$$

Table 12.4.1 Conditional means and SDs of weight given height in a population of young men*

Height (in) X	Mean weight (lb) $\mu_{Y\|X}$	Standard deviation of weights (lb) $\sigma_{Y\|X}$
64	127	20
68	144	20
72	161	20
76	178	20

*Note that all values of $\sigma_{Y|X}$ are the same; they equal $\sigma_\varepsilon = 20$.

Weight vs. height



Estimating β_0, β_1, and σ_ϵ

- b_0 estimates β_0.
- b_1 estimates β_1.
- s_e estimates σ_e.
- Example 12.4.5. For the snake data, $b_0 = -301$ estimates β_0, $b_1 = 7.19$ estimates β_1, and $s_e = 12.5$ estimates σ_e.
- We estimate the the mean weight $\hat{y}$ of snakes with length x as

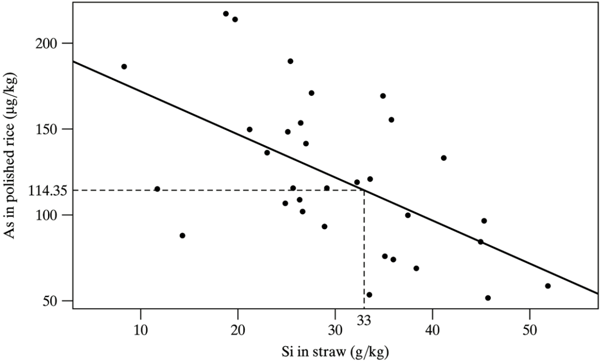
$$\hat{y} = -301 + 7.19x$$

Example 12.4.6 Arsenic in rice

- If we believe the data follow a line, we can estimate the mean for any x we want.
- $b_0 = 197.17$ estimates β_0, $b_1 = 2.51$ estimates β_1, and $s_e = 37.30$ estimates σ_e.
- For straw silicon concentration of $x = 33$ g/kg we estimate a mean arsenic level of

$$\hat{y} = 197.17 - 2.51(33) = 114.35 \ \mu\text{gm/kg with } s_e = 37.30 \ \mu\text{gm/kg}.$$

Arsenic in rice at $X = 33$ g/kg



$$\hat{y} = 197.17 - 2.51x$$

$$114.35 = 197.17 - 2.51(33)$$

12.5 Inference for β_1

- Often people want a 95% confidence interval for β_1 and want to test $H_0 : \beta_1 = 0$.
- If we reject $H_0 : \beta_1 = 0$, then y is significantly linearly assocatied with x. Same as testing $H_0 : \rho = 0$.
- A 95% confidence interval for β_1 gives us a range for how the mean changes when x is increased by one unit.
- Everything comes from

$$\frac{b_1 - \beta_0}{SE_{b_1}} \sim t_{n-2}, \quad SE_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}.$$

- R automatically gives a P-value for testing $H_0 : \beta_1 = 0$.
- Need to ask R for 95% confidence interval for β_1.

R code

```
> amph=c(0,0,0,0,0,0,0,0,2.5,2.5,2.5,2.5,2.5,2.5,2.5,2.5,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0)
> cons=c(112.6,102.1,90.2,81.5,105.6,93.0,106.6,108.3,73.3,84.8,67.3,55.3,
+        80.7,90.0,75.5,77.1,38.5,81.3,57.1,62.3,51.5,48.3,42.7,57.9)
> fit=lm(cons~amph)
> summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   99.331      3.680   26.99  < 2e-16 ***
amph          -9.007      1.140   -7.90 7.27e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> confint(fit)
                2.5 %     97.5 %
(Intercept)  91.69979 106.962710
amph        -11.37202  -6.642979
```

P-value for testing $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ is 0.0000000727, we reject at the 5% level. We are 95% confidence that true mean consumption is reduced by 6.6 to 11.4 g/kg for every mg/kg increase in amphetamine dose.

Multiple regression

- Often there are more than one predictors we are interested in, say we have two x_1 and x_2.
- The model is easily extended to

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- Example: Dwayne Portrait Studio is doing a sales analysis based on data from $n = 21$ cities.
  - $y =$ sales (thousands of dollars) for a city
  - $x_1 =$ number of people 16 years or younger (thousands)
  - $x_2 =$ per capita disposable income (thousands of dollars)

The data

x_1	x_2	y	x_1	x_2	y
68.5	16.7	174.4	45.2	16.8	164.4
91.3	18.2	244.2	47.8	16.3	154.6
46.9	17.3	181.6	66.1	18.2	207.5
49.5	15.9	152.8	52.0	17.2	163.2
48.9	16.6	145.4	38.4	16.0	137.2
87.9	18.3	241.9	72.8	17.1	191.1
88.4	17.4	232.0	42.9	15.8	145.3
52.5	17.8	161.1	85.7	18.4	209.7
41.3	16.5	146.4	51.7	16.3	144.0
89.6	18.1	232.6	82.7	19.1	224.1
52.3	16.0	166.5			

R code for multiple regression

```
> under16=c(68.5,45.2,91.3,47.8,46.9,66.1,49.5,52.0,48.9,38.4,87.9,72.8,88.4,42.9,52.5,
+          85.7,41.3,51.7,89.6,82.7,52.3)
>
> income=c(16.7,16.8,18.2,16.3,17.3,18.2,15.9,17.2,16.6,16.0,18.3,17.1,17.4,15.8,17.8,
+         18.4,16.5,16.3,18.1,19.1,16.0)
>
> sales=c(174.4,164.4,244.2,154.6,181.6,207.5,152.8,163.2,145.4,137.2,241.9,191.1,232.0,
+        145.3,161.1,209.7,146.4,144.0,232.6,224.1,166.5)
> fit=lm(sales~under16+income)
> summary(fit)

Call:
lm(formula = sales ~ under16 + income)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4239  -6.2161   0.7449   9.4356  20.2151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
under16       1.4546     0.2118   6.868    2e-06 ***
income        9.3655     4.0640   2.305   0.0333 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared: 0.9167,     Adjusted R-squared: 0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

Interpretation...

- The fitted regression *surface* is

$$\text{sales} = -68.857 + 1.455\ (\text{under } 16) + 9.366\ \text{income}.$$

- For every unit increase (1000 people) in those under 16, average sales go up 1.455 thousand, \$1,455.
- For every unit increase (\$1000) in disposable income, average sales go up 9.366 thousand, \$9,366.
- 91.67% of the variability in sales is explained by those under 16 and disposable income.
- σ_e is estimated to be 11.01.

Regression homework

- 12.2.5, 12.2.7, 12.3.1, 12.3.3, 12.3.5, 12.3.7, 12.3.8. Use R for all problems; i.e. don't do anything by hand.
- 12.4.3, 12.4.6, 12.4.8, 12.4.9, 12.5.1, 12.5.3, 12.5.5, 12.5.9(a). Use R for all problems; don't do anything by hand.