* Announcement :

① Quiz 11 : Thu (12/2) 9:00 AM ~ Fri (12/3) 9:00 AM

"30 minutes"

Ch11.2 ~ Ch12.2

- Three estimators : $\begin{cases} T_1 \\ T_2 \\ T_3 \end{cases}$ to estimate the parameter from uniform dist'n (Ch11.2)
- Least squares regression. (Ch11.3)
- Correlation (Ch11.4) : $r(aX+b, cY+d) = \begin{cases} r(X,Y) & ac>0 \\ -r(X,Y) & ac<0 \end{cases}$
- Exp, var of residual $D = Y - \hat{Y}$, other properties (Ch11.5)
- Review Exercise 11.6.8 / 11.6.11

STAT 88: Lecture 39

- Assumptions of simple linear regression model, Dist'n of $\hat{\beta}_1$, T, etc (Ch12.1, 12.2)

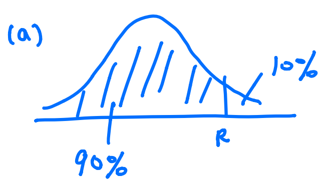
Contents

Warm up: (Related to Exercise 11.6.8) Assume R and S are normal. The correlation between R and S is 0.6, i.e. $r(R, S) = 0.6$.

(a) If R is 90th percentile, estimate the percentile rank of S.

(b) If R is 10th percentile, estimate the percentile rank of S.

(a)



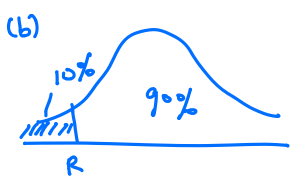
$\Phi\left(\frac{R-\mu_R}{\sigma_R}\right) = \Phi(R^*) = 0.9$

$\Rightarrow R^* = \Phi^{-1}(0.9)$

$\Rightarrow \hat{S}^* = r \cdot R^* = 0.6 \cdot \Phi^{-1}(0.9)$ (Lecture 36, Page 5)

$\Rightarrow \Phi(\hat{S}^*) = \Phi(0.6 \cdot \Phi^{-1}(0.9))$

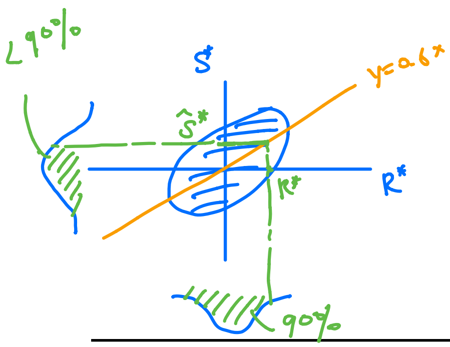
(b)



$R^* = \Phi^{-1}(0.1)$

$\Rightarrow \hat{S}^* = r \cdot R^* = 0.6 \cdot \Phi^{-1}(0.1)$

$\Rightarrow \Phi(\hat{S}^*) = \Phi(0.6 \cdot \Phi^{-1}(0.1))$



$\begin{cases} R > 50\text{th percentile} \Rightarrow \text{Percentile of } R > \text{" of } S \\ R < 50\text{th percentile} \Rightarrow \text{"} < \text{"} \end{cases}$

$(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$

$\to \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Last time

The distribution of the estimated slope

$$\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

$D_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

σ is unknown so we estimate it with the SD of the residuals. Since

$$\mathrm{SD}(\widehat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

"constant"

we have

SD of the residuals

$$\mathrm{SE}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

"random"

where $\widehat{\sigma}$ is the SD of residuals. Therefore, when n is large,

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{SE}(\widehat{\beta}_1)} \sim \mathcal{N}(0, 1).$$

Approximate (only when n is large)

12.2. The Distribution of the Estimated Slope

Pulse Rates

We wish to predict active pulse rates from resting pulse rates.

```
pulse
```

Active	Rest	Smoke	Sex	Exercise	Hgt	Wgt
97	78	0	1	1	63	119
82	68	1	0	3	70	225
88	62	0	0	3	72	175
106	74	0	0	3	72	170
78	63	0	1	3	67	125
109	65	0	0	3	74	188
66	43	0	1	3	67	140
68	65	0	0	3	70	200
100	63	0	0	1	70	165
70	59	0	1	2	65	115

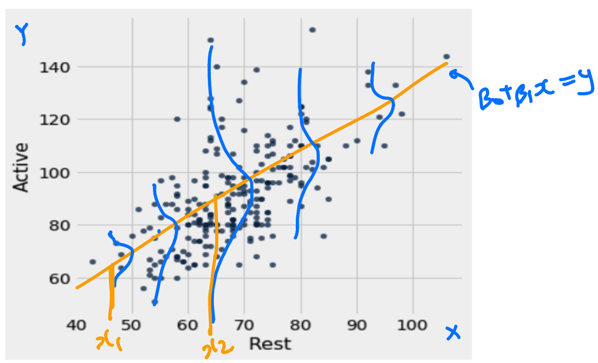
... (222 rows omitted)

```
pulse.scatter('Rest', 'Active')
```

Assumptions of simple linear regression model?

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Constant variance assumption is violated.



$\sigma_1^2 \ll \sigma_2^2$

```
active = pulse.column(0)     Y
resting = pulse.column(1)    X
```

```
stats.linregress(x=resting, y=active)
```

```
(1.142879681904831,        β̂1
 13.182572776013345,       β̂0
 0.6041870881060092,       r
 1.7861044071652305e-24,   p-val
 0.09938884436389145)      SE(β̂1)
```

$H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$

$n = 232$ is large so

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{SE}(\widehat{\beta}_1)} \overset{\text{Approximately}}{\sim} \mathcal{N}(0, 1).$$

$$\underbrace{P\left(-2 \le \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{SE}(\widehat{\beta}_1)} \le 2\right)}_{\beta_1 \in [\widehat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\widehat{\beta}_1)]} = 95\%$$

A 95% CI for β_1 is

$$(\widehat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\widehat{\beta}_1)) = (0.944, 1.342).$$

A fundamentally important question is whether the true slope β_1 is 0. If it is 0, then the resting pulse rate isn't involved in the prediction of the active pulse rate, according to the regression model. Our testing problem is

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0.$$

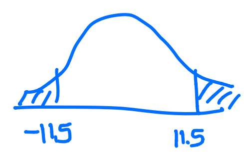
T is our test statistic. Under H_0,

$$\underbrace{T}_{\beta_1 = 0} = \frac{\widehat{\beta}_1}{\mathrm{SE}(\widehat{\beta}_1)} \sim \mathcal{N}(0, 1).$$

$$= \frac{1.1428}{0.09938}$$

The observed value of the test statistic is 11.5. So the p-value is

$$\text{p-value} = P(T \geq 11.5) + P(T \leq -11.5) \approx 0.$$



We reject H_0 at 5% level.

t Statistic

Above we assume that n is large so

$$= \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$\mathrm{SE}(\widehat{\beta}_1) \approx \mathrm{SD}(\widehat{\beta}_1).$$

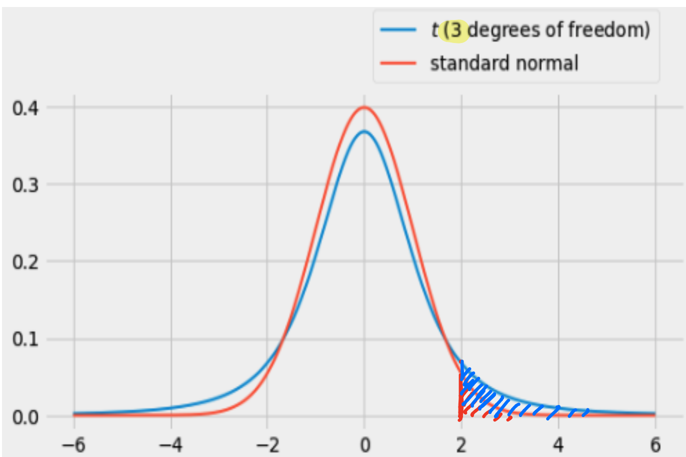
$$= \frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)}$$

If n is small, this approximation is not good and T has a *t-distribution* with $n-2$ as a parameter (called degrees of freedom).

$\begin{cases} N(\mu,\sigma^2) \\ t(df) \end{cases}$

t-distribution: The family of t-distributions is indexed by the positive integers: there's the t-distribution(1), the t-distribution(2), and so on.

The t density looks like the standard normal curve, except that it has fatter tails.



$P(Z>2)$

$P(\cancel{|Z|>2})$ for $Z \sim N(0,1)$

$\|$

- 1 − stats.norm.cdf(2)

  → 0.022750

parameter

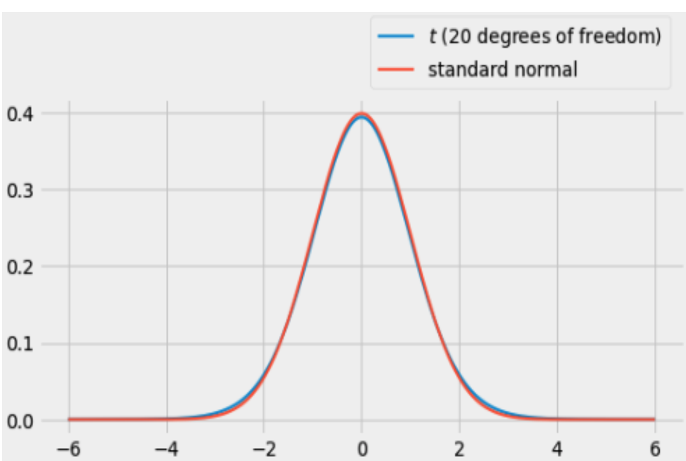
- 1 − stats.t.cdf(2, df=3)

  → 0.069662

$\|$

$P(\cancel{|T|>2})$ for $T \sim t(3)$

$P(T>2)$



"fact"

$t(df) \to N(0,1)$ as $df \to \infty$

$SD(D_1, \dots, D_n)$, $D_i = Y_i - \hat{Y}_i = Y_i - \hat\beta_0 - \hat\beta_1 x_i$

$SE = \dfrac{\hat\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$

exact dist'n

n small $\Rightarrow$ use $T \sim t(n-2)$

n large $\Rightarrow$ use $T \sim N(0,1)$

FACT:

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{SE}(\widehat{\beta}_1)} \sim t(n-2).$$

n large $t(n-2) \approx N(0,1)$

The n is because there are n independent observations and the -2 is because there are two parameter estimates we need to make.

β_1, β_0

Example: (Exercise 12.4.3) Refer to the regression of active pulse rate on resting pulse rate in Section 12.2. Here are the estimated values again, along with some additional data.

$\hat\beta_1$, $\hat\beta_0$, r, p-val, $SE(\hat\beta_1)$

```
(1.142879681904831,
 13.182572776013345,
 0.6041870881060092,
 1.7861044071652305e-24,
 0.09938884436389145)
```

```
mean_active, sd_active = np.mean(active), np.std(active)
mean_active, sd_active
```

$= \bar{Y}$, $= SD(Y_1, \dots, Y_n)$

```
(91.29741379310344, 18.779629284683832)
```

```
mean_resting, sd_resting = np.mean(resting), np.std(resting)
mean_resting, sd_resting
```

$= \bar{x}$, $= SD(x_1, \dots, x_n)$

```
(68.34913793103448, 9.927912546587986)
```

c) Find the SD of the residuals.

$\hat\sigma$ = SD of residuals.

$$SE(\hat\beta_1) = \frac{\hat\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\Rightarrow \hat\sigma = \underbrace{SE(\hat\beta_1)}_{= 0.09938} \cdot \underbrace{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\text{sd-resting} = \sqrt{\tfrac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = 9.9279$$

$$\Rightarrow \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{n} \cdot 9.9279 = \sqrt{232} \cdot 9.9279$$

$$= 0.09938 \cdot \sqrt{232} \cdot 9.9279$$

$$\approx 14.97$$

Example: Restricting the pulse regression data to male smokers. The sample size reduces $n = 17$.

You get the following readout:

$SE(\hat\beta)$ $\quad \frac{\hat\beta - \beta}{SE(\hat\beta)}$ $\quad$ p-value $\quad$ 95% CI for β

		coef	std err	t	P>|t|	[0.025	0.975]
β_0	**const**	9.9360	16.345	0.608	0.552	-24.903	44.775
β_1	**Rest**	1.1591	0.222	5.224	0.000	0.686	1.632

$\hat\beta_0$ $\quad \hat\beta_1$ $\quad SE(\hat\beta_1)$ $\quad \frac{\hat\beta_1 - \beta_1}{SE(\hat\beta_1)}$

What can you conclude from this?

$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

$\Rightarrow$ p-value ≈ 0

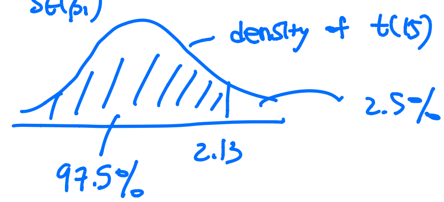
$\Rightarrow$ Reject H_0 at 5% level.

Given

```
stats.t.ppf(.975,df=15)
```

```
2.131449545559323
```

$$T = \frac{\hat\beta_1 - \beta_1}{SE(\hat\beta_1)} \sim t(15)$$



Verify the 95% CI for β_1 is $[0.686, 1.632]$.

$$\Rightarrow P\left(-2.13 \le \frac{\hat\beta_1 - \beta_1}{SE(\hat\beta_1)} \le 2.13\right) = 95\%$$

$$\Updownarrow$$

$$\beta_1 \in [\hat\beta_1 \pm 2.13 \cdot SE(\hat\beta_1)]$$

$\hat\beta_1 \pm 2.13 \cdot SE(\hat\beta_1)$

$= 1.1591 \pm 2.13 \cdot 0.222$

$= [0.686, 1.632]$

12.3. Towards Multiple Regression

Below is data on a random sample of Hodgkin cancer patients.

Simple Regression

We predict difference from base:

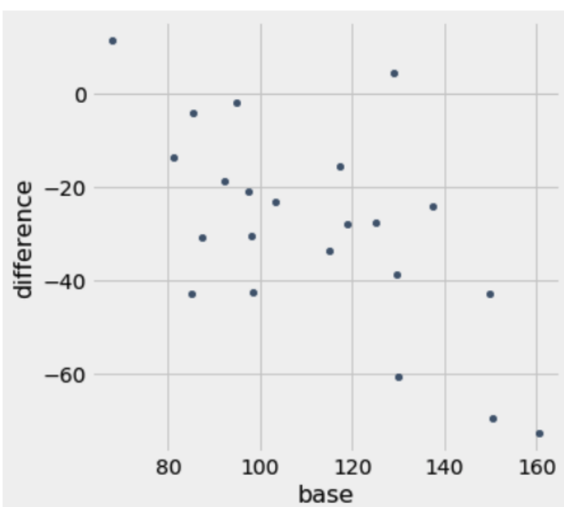
```
hodgkins
```

Health before chemo (bigger means more healthy)

height	rad	chemo	base	month15	difference
164	679	180	160.57	87.77	-72.8
168	311	180	98.24	67.62	-30.62
173	388	239	129.04	133.33	4.29
157	370	168	85.41	81.28	-4.13
160	468	151	67.94	79.26	11.32
170	341	96	150.51	80.97	-69.54
163	453	134	129.88	69.24	-60.64
175	529	264	87.45	56.48	-30.97
185	392	240	149.84	106.99	-42.85
178	479	216	92.24	73.43	-18.81

... (12 rows omitted)

```
hodgkins.scatter('base', 'difference')
```



n = 22

OLS Regression Results

Dep. Variable:	difference	**R-squared:**	0.397

	coef	std err	t	P>|t|	[0.025	0.975]
const	32.1721	17.151	1.876	0.075	-3.604	67.949
base	-0.5447	0.150	-3.630	0.002	-0.858	-0.232

What difference do you predict if you have base health 100?

Multiple Regression

What if we want to regress on both base and chemo? Here chemo is very uncorrelated with base.

```
h_data.corr()
```

	height	rad	chemo	base	month1!
height	1.000000	-0.305206	0.576825	0.354229	0.39052
rad	-0.305206	1.000000	-0.003739	0.096432	0.04061
chemo	0.576825	-0.003739	1.000000	0.062187	0.44578
base	0.354229	0.096432	0.062187	1.000000	0.56137
month15	0.390527	0.040616	0.445788	0.561371	1.00000
difference	-0.043394	-0.073453	0.346310	-0.630183	0.28879

Conceptual picture:

$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

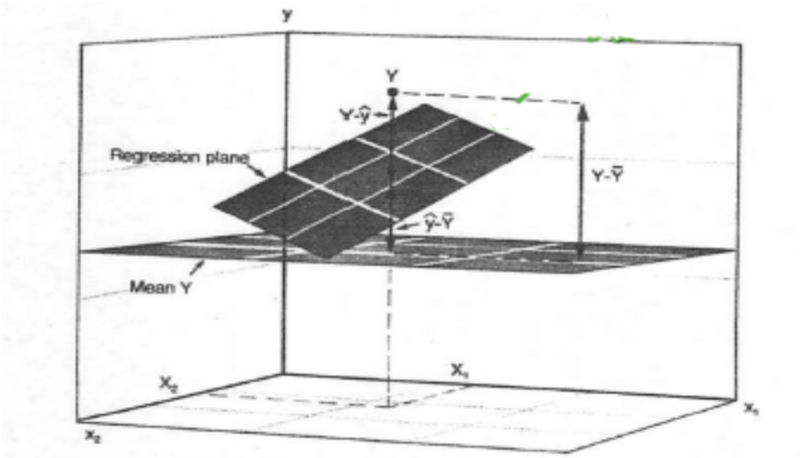


FIGURE 3-5 The deviation of the observed value of Y from the mean of all values of Y, $(Y - \bar{Y})$, can be separated into two components: the deviation of the observed value of Y from the value on the regression plane $(Y - \hat{y})$ at the associated values of the independent variables X_1 and X_2, and the deviation of the regression plane from the observed mean value of $\bar{Y}$ $(\hat{y} - \bar{Y})$ (compare with Fig. 2-7).

OLS Regression Results

Dep. Variable:	difference	R-squared:	0.546

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.9992	20.227	-0.049	0.961	-43.335	41.336
base	-0.5655	0.134	-4.226	0.000	-0.846	-0.285
chemo	0.1898	0.076	2.500	0.022	0.031	0.349

What can you conclude here about the fit and $\beta_0, \beta_1, \beta_2$?

What if we include all features?

```
h_data.corr()
```

	height	rad	chemo	base	month1
height	1.000000	-0.305206	0.576825	0.354229	0.39052
rad	-0.305206	1.000000	-0.003739	0.096432	0.04061
chemo	0.576825	-0.003739	1.000000	0.062187	0.44578
base	0.354229	0.096432	0.062187	1.000000	0.56137
month15	0.390527	0.040616	0.445788	0.561371	1.00000
difference	-0.043394	-0.073453	0.346310	-0.630183	0.28879

Note that we have multi-collinearity (i.e. some features are highly correlated with each other).

OLS Regression Results

Dep. Variable:	difference	**R-squared:**	0.550

a very minor improvement

	coef	std err	t	P>|t|	[0.025	0.975]
const	33.5226	101.061	0.332	0.744	-179.698	246.743
base	-0.5393	0.160	-3.378	0.004	-0.876	-0.202
chemo	0.2124	0.103	2.053	0.056	-0.006	0.431
rad	-0.0062	0.031	-0.203	0.841	-0.071	0.059
height	-0.2274	0.658	-0.346	0.734	-1.615	1.160