Stat 88 lec 39

Warmup 2:00–2:10

Suppose you are doing linear regression to predict active pulse from resting pulse.
You have the following Python output:

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

What can you conclude from this?

The first row:

$H_0: \beta_0 = 0$ ← at zero resting HR you have 0 active HR.

$H_A: \beta_0 \neq 0$

this has p-val .552 > .05

so accept the null.

The second row:

$H_0: \beta_1 = 0$ ← resting and active HR are independent

$H_A: \beta_1 \neq 0$

this p-value $\approx 0 < .05$

so reject the null.

Future lectures

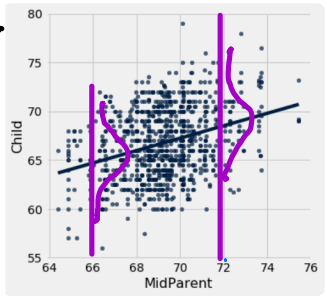
today → W finish
F review

RRR week W review

Last time

Sec 11.5 The error in regression

let $D = \hat{y} - y$ be the residual error



$$D \sim N\left(0, (1-r^2)\sigma_y^2\right)$$

$E(D)$ $\quad$ $Var(D)$

Sec 12.2 The distribution of the estimated slope.

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{n\,var(x)}\right)$$

Hence $SD(\hat{\beta}_1) = \dfrac{\sigma}{\sqrt{n\,var(x)}}$ ← unknown parameter

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n\,var(x)}} \quad \text{where } \hat{\sigma} = \text{SD of residuals}$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$$

for large n

is our test statistic for

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$.

Today ① Sec 12.2 The distribution of the estimated slope

② Sec 12.3 towards multiple regression

① Sec 12.2 the distribution of the estimated slope.

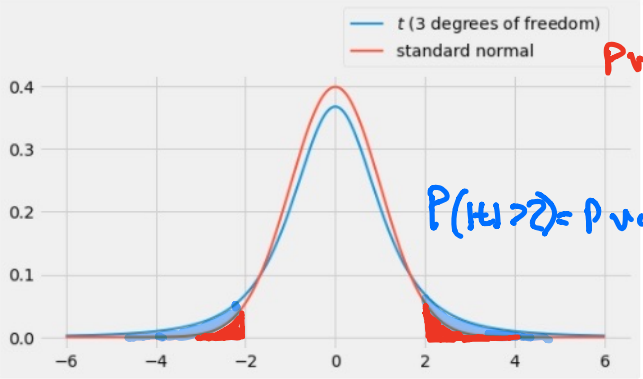
t-statistic:

Above we assume that n is large so

$SE(\hat{B}_1) \approx SD(\hat{B}_1)$

for small n, T has a t-distribution with $n-2$ as parameters (called degrees of freedom)

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

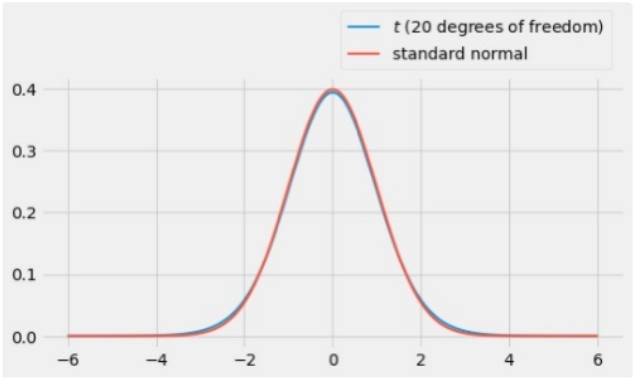


Pval = `1-stats.norm.cdf(2)`

```
0.02275013194817921
```

$P(|t| > 2) = Pval =$ `1-stats.t.cdf(2,df=3)`

```
0.06966298427942152
```



for n large you can approximate the t and std normal distributions.

Example 12.4.3

In the pulse example from last class

$\hat{\beta}_1$
$\hat{\beta}_0$
r
p-val
$SE(\hat{\beta}_1)$

```
(1.142879681904831,
 13.182572776013345,
 0.6041870881060092,
 1.7861044071652305e-24,
 0.0993888436389145)
```

μ_y = σ_y

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

```
(91.29741379310344, 18.779629284683832)
```

μ_x = σ_x

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

```
(68.34913793103448, 9.927912546587986)
```

c) Find the SD of the residuals.

Two ways to do this:

$r = .604$

$\sigma_y = 18.78$

$SD(R) = \sqrt{1-r^2}\,\sigma_y$

$n = 232$

So $SD(R) = \sqrt{1-r^2}\,\sigma_y = \sqrt{1-(.604)^2}\,(18.78) \approx \boxed{14.96}$

OR

Calculate $\hat{\sigma} = SD(R)$ from $SE(\hat{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{n\sigma_x^2}}$

$\Rightarrow \hat{\sigma} = SE(\hat{\beta}_1)\sqrt{n\sigma_x^2}$

$= (.0993)\sqrt{232\,(9.93)} = \boxed{14.99}$

Ex restricting the pulse regression data to male nonsmokers the sample size reduces to n=17.

You get the following readout:

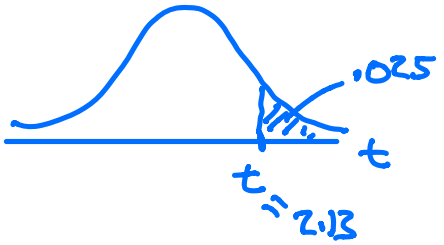
	coef	std err	t	P>|t|	[0.025	0.975]
const	9.9360	16.345	0.608	0.552	-24.903	44.775
Rest	1.1591	0.222	5.224	0.000	0.686	1.632

$\hat{\beta}_0$ $\hat{\beta}_1$

Given,

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

```
2.131449545559323
```



Verify the 95% CI for β_1 is [.686, 1.632].

95% CI for β_1 is $\hat{\beta}_1 \pm t_{n-2}(.025)\,SE(\hat{\beta}_1)$ (2.13, .222)

$= 1.16 \pm 2.13(.222) = [.687, 1.633]$

What can you conclude regarding your hypothesis test?

⇒ We reject the null that $\beta_1 = 0$ for the alternative $\beta_1 \neq 0$ at level .05.

② Sec 12.3 towards multiple regression

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

simple regression

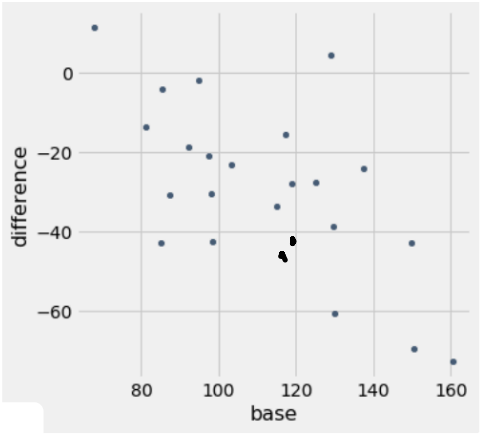
we predict difference from base

```
h_data = hodgkins.to_df()
h_data
```

health before chemo (bigger means more healthy)

	height	rad	chemo	base	month15	difference
0	164	679	180	160.57	87.77	-72.80
1	168	311	180	98.24	67.62	-30.62
2	173	388	239	129.04	133.33	4.29
3	157	370	168	85.41	81.28	-4.13
4	160	468	151	67.94	79.26	11.32

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



n = 27

r^2

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?

$\hat{\beta}_0 + \hat{\beta}_1(100)$

$= 32.17 + (-.5447)(100) = -22.3$

multiple regression

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

```
h_data.corr()
```

	height	rad	chemo	base	month
height	1.000000	-0.305206	0.576825	0.354229	0.3905
rad	-0.305206	1.000000	-0.003739	0.096432	0.0406
chemo	0.576825	-0.003739	1.000000	0.062187	0.4457
base	0.354229	0.096432	0.062187	1.000000	0.5613
month15	0.390527	0.040616	0.445788	0.561371	1.0000
difference	-0.043394	-0.073453	0.346310	-0.630183	0.2887

Conceptual picture:

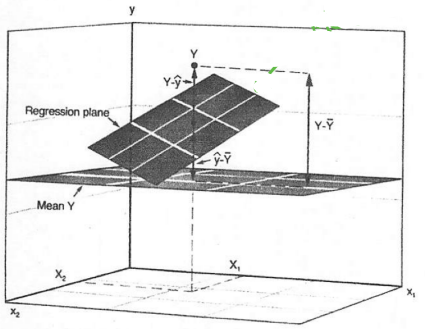


FIGURE 3-5 The deviation of the observed value of Y from the mean of all values of Y, $(Y - \bar{Y})$, can be separated into two components: the deviation of the observed value of Y from the value on the regression plane $(Y - \hat{y})$ at the associated values of the independent variables X_1 and X_2,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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

sults

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 $\beta_0, \beta_1, \beta_2$?

Higher R^2 so better fit,

Conclude β_0 not significantly diff than 0

and β_1 and β_2 are significantly diff than zero.

What if we include all predictors?

```
h_data.corr()
```

	height	rad	chemo	base	month
height	1.000000	-0.305206	0.576825	0.354229	0.3905
rad	-0.305206	1.000000	-0.003739	0.096432	0.0406
chemo	0.576825	-0.003739	1.000000	0.062187	0.4457
base	0.354229	0.096432	0.062187	1.000000	0.5613
month15	0.390527	0.040616	0.445788	0.561371	1.0000
difference	-0.043394	-0.073453	0.346310	-0.630183	0.2887

Note that we have multicollinearity (i.e some predictors are correlated with each other).

OLS Regression Results

Dep. Variable:	difference	R-squared:	0.550

a very minor improvement mostly from adding rad

	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

Notice we have multicollinearity of some of the predictors (i.e height and month15 are related to chemo and base so they dont do much for you),

So it is best just to have 2 predictors base and chemo.