Stat 88 lec 40

Today ① Review chap 12 (simple and multiple regression)
② overview of class
③ method of indicators

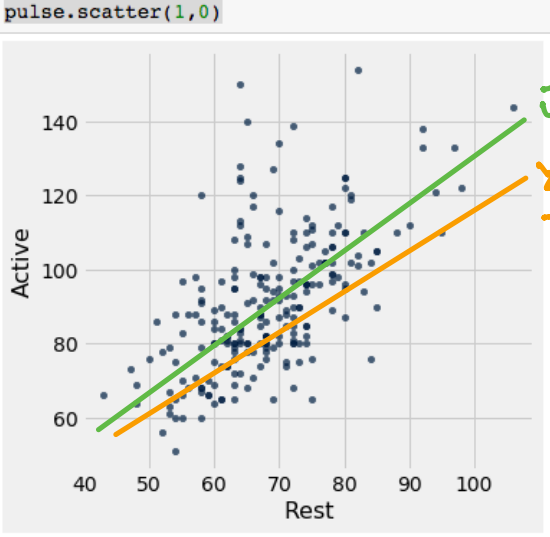
Ex

```
pulse
```

Active	Rest
97	78
82	68
88	62
106	74
78	63
109	65
66	43
68	65
100	63
70	59

... (222 rows omitted) $(n=232)$

```
pulse.scatter(1,0)
```



regression line $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$

$y=\beta_0+\beta_1 x$ Tyche line

To test the hypothesis

$H_0: \beta_1 = 0$

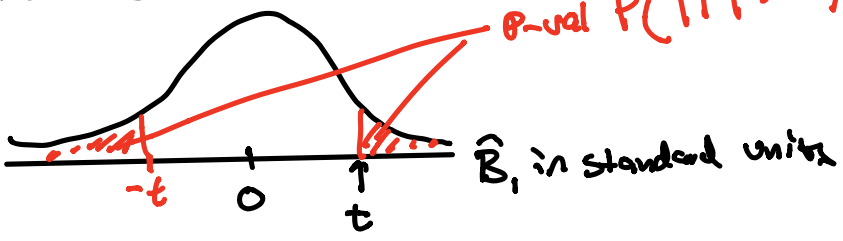
$H_A: \beta_1 \neq 0$

use T.S.

$$T = \frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{n}\sigma_x}} \quad \text{where} \quad \hat{\sigma} = \sqrt{1-r^2}\,\sigma_y$$

$\frac{\hat{\sigma}}{\sqrt{n}\sigma_x} = SE(\hat{\beta})$

Assuming the null



OLS Regression Results

Dep. Variable:	Active	R-squared:	0.365

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.1826	6.864	1.920	0.056	-0.343	26.708
Rest	1.1429	0.099	11.499	0.000	0.947	1.339

r^2, p-val, 95% CI, SE, T, β_0, $\hat{\beta}_0$, β_1, $\hat{\beta}_1$

Which can you conclude from this table about β_0, β_1? — we accept null that $\beta_0 = 0$ and reject null that $\beta_1 = 0$

If I dont give you t in this table can you figure it out from the rest of the table?

↳ yes, assuming the null

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{1.1429}{0.099} = 11.499$$

If I dont give you P>|t| in this table can you figure it out from the rest of the table?

↳ we know t

$$\text{p-val} = P>|t| = 2(1-\Phi(t))$$ ← for n large

since $n=232$ is large and $T \sim N(0,1)$

or $2(1-\text{stats.t.cdf}(t, df=n-2))$

OLS Regression Results

Dep. Variable:	Active	R-squared:	0.365

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.1826	6.864	1.920	0.056		
Rest	1.1429	0.099	11.499	0.000		

Can you find the 95% CI for β_1 from the table above?

↳ yes $\hat{\beta}_1 \pm 2\,SE(\hat{\beta}_1) = 13.18 \pm 2(6.864)$

↑ or use stats.t.ppf(.975, df=n-2) if $n<30$

OLS Regression Results

Dep. Variable:	Active	R-squared:	0.365

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.1826	6.864	1.920	0.056	-0.343	26.708
Rest	1.1429	0.099	11.499	0.000	0.947	1.339

$\mu_y = 91.29$

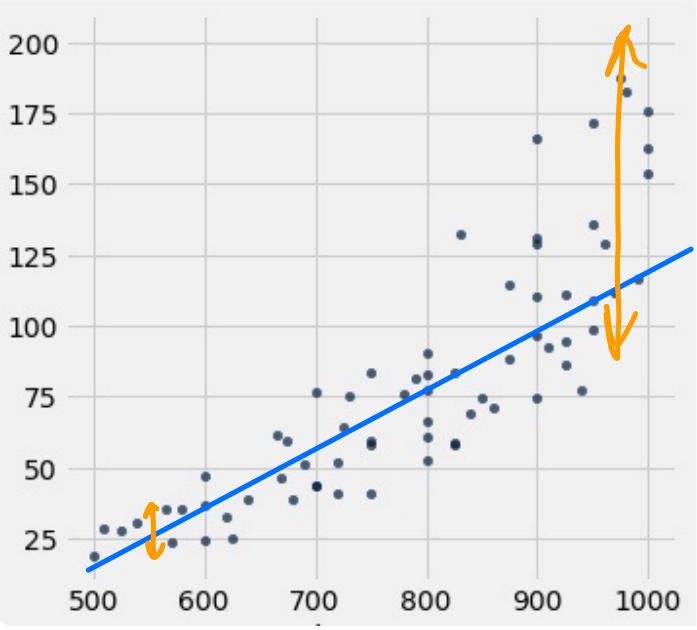
$\sigma_y = 18.78$

If I give you the info above can you find μ_x and σ_x?

$\hat{\beta}_1 = r\frac{\sigma_y}{\sigma_x} \Rightarrow \sigma_x = \frac{r\sigma_y}{\hat{\beta}_1} = \frac{\sqrt{.365}\,(18.78)}{1.143} = \boxed{9.97}$

$\hat{\beta}_0 = \mu_y - \hat{\beta}_1\mu_x \Rightarrow \mu_x = \frac{\mu_y - \hat{\beta}_0}{\hat{\beta}_1} = \frac{91.29 - 13.18}{1.14} = \boxed{68.52}$

ex Suppose you have a scatter diagram given below, should you make a linear regression model? Explain.



No our linear model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$ the errors here are not having a constant variance (i.e variance bigger for bigger x).

Sec 12.3 Multiple regression:

multiple regression generalizes simple linear regression to more than one predictor.

ex Now Rest and Wgt will be predictors. They should not be correlated ($r = -.18$) ok

pulse

Active	Rest	Wgt
97	78	119
82	68	225
88	62	175
106	74	170
78	63	125
109	65	188
66	43	140
68	65	200
100	63	165
70	59	115

... (222 rows omitted)

	Active	Rest	Wgt
Active	1.000000	0.604187	-0.058012
Rest	0.604187	1.000000	-0.183928
Wgt	-0.058012	-0.183928	1.000000

Dep. Variable:			Active		R-squared:		0.368
		coef	std err	t	P>|t|	[0.025	0.975]
β_0	const	6.7425	9.290	0.726	0.469	-11.563	25.048
β_1	Rest	1.1620	0.101	11.494	0.000	0.963	1.361
β_2	Wgt	0.0325	0.032	1.029	0.305	-0.030	0.095

What can you conclude from this table?

accept null for β_0, β_2 and reject for β_1

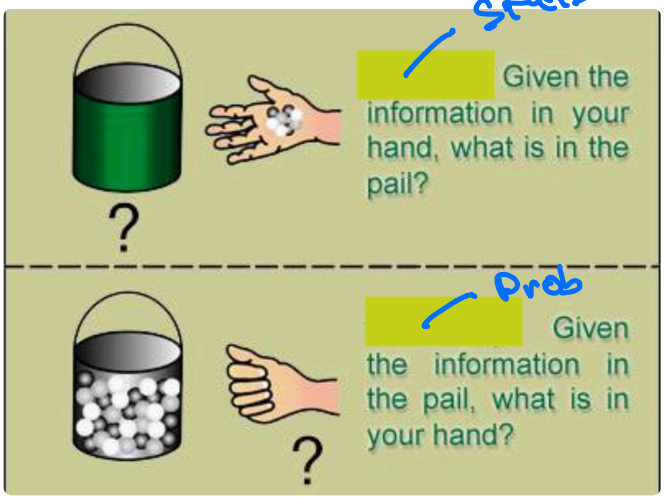
R^2 goes up just a little so wgt isn't contributing to a better fit. We should remove wgt from the model since $\beta_2 = 0$.

What Active pulse would you expect if you have if Rest = 50 and Wgt = 150?

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \text{rest} + \hat{\beta}_2 \text{wgt}$

$= 6.74 + 1.16(50) + .03(150) = \boxed{69.24}$

Overview of class



Stats = "inverse prob"

One first day of class I asked you which is probability and which is statistics.

Probability

You have learned both discrete and continuous probability distributions

discrete

Name and Parameters
Bernoulli (p) (also Indicator)
Uniform on 1 through N
Binomial (n, p)
Hypergeometric (N, G, n)
Poisson (μ)
Geometric (p)

continuous

Name and Parameters
Uniform (a, b) ← analogous to Unif on 1 through N
Exponential (λ) ← analogous to Geom (p)
Normal (μ, σ^2) ← analogous to Binomial (n, p)

You have learned how to calculate the expectation of quite complicated probability distributions using the method of indicators. Expectation is the center or average of your data's histogram.

You learned how that SD is the average spread of your data from the mean.

If we dont assume anything about the population distribution except the mean and SD you can use Chebyshev inequality to get an upperbound on the tail probability.

Large sample approximations such as the law of averages which says $\bar{X} \to \mu_X$ and the CLT which says $\bar{X} \sim N(\mu_X, \sigma_X^2)$ is very useful.

Inference (statistics)

Given a sample from a population with one of the above distributions you learned how to estimate the parameter population.

In the case of regression you learned how to use the regression line to estimate the Tyche line.

From our sample we can make hypotheses about the value of the parameter of the population distribution. Assuming the null is true we compute a test statistic and compute the p-value. If the p-value is less that .05 (for a .05 level test) we reject the null.

A 95% CI for an unknown parameter tells you whether you should reject the null for the alternative.

(3) Method of Indicators

In a box of tickets, 60% of the tickets are blue, 20% green, 15% yellow, and 5% purple. Find the expected number of colors that do not appear among d draws made at random ~~with~~ without replacement from the box.

Since the population size N isn't given you can assume it is infinite, so drawing with and without replacement are equivalent,

The solution then is $E(X) = (.4)^d + (.8)^d + (.85)^d + (.95)^d$, $d \geq 4$

(see midterm review for details)

If the problem tells you N=100 then the answer is

X = the # of colors (out of 4) that do not appear

$I_G = \begin{cases} 1 & \text{if green doesn't appear in d draws} \\ 0 & \text{else} \end{cases}$ (G = green, P_G)

Similarly for other colors

$$E(X) = \underbrace{\frac{\binom{60}{0}\binom{40}{d}}{\binom{100}{d}}}_{\text{Blue}} + \underbrace{\frac{\binom{20}{0}\binom{80}{d}}{\binom{100}{d}}}_{\text{Green}} + \underbrace{\frac{\binom{15}{0}\binom{85}{d}}{\binom{100}{d}}}_{\text{Yellow}} + \underbrace{\frac{\binom{5}{0}\binom{95}{d}}{\binom{100}{d}}}_{\text{purple}}$$

(P_G = // Green)