* Announcement:
① HW13 due 12/7
② Final review schedule during RRR week
post on Piazza soon
③ OH during RRR week proceed as usual

STAT 88: Lecture 40

Contents

Warm up: You get the following readout for the simple linear regression model:

$n=232$

	coef ($\hat{\beta}$)	std err ($SE(\hat{\beta})$)	t ($\frac{\hat{\beta}}{SE(\hat{\beta})}$)	P>|t| (p-value)	[0.025 (95% CI)	0.975]
β_0 const	13.1826	6.864	1.920	0.056		
β_1 Rest	1.1429	0.099	11.499	0.000		

0.056 → p-value for $H_0: \beta_0=0$ vs $H_1: \beta_0 \neq 0$

What can you conclude from this table about β_1?

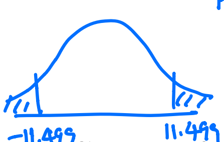
p-val ≈ 0 $\Rightarrow$ $H_0: \beta_1=0$ vs $\beta_1 \neq 0$.
Reject H_0 at level 5%

If I don't give you t in this table, can you figure it out from the rest of the table?

Yes, $T=\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{1.1429}{0.099} = 11.499$

If I don't give you $P > |t|$ in this table, can you figure it out from the rest of the table?

$$\text{p-val} = P(T>11.499) + \underbrace{P(T<-11.499)}_{=\Phi(-11.499)} = \begin{cases} 2(1-\Phi(11.499)) & \leftarrow n \text{ is large and } T \sim N(0,1) \\ 2(1-\text{stats.t.cdf}(11.499, df=n-2)) & \leftarrow n \text{ is small and } t \sim t(df=n-2) \end{cases}$$



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

$$\begin{cases} \hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1) & \text{if } n \text{ is large} \\ \hat{\beta}_1 \pm \text{stats.t.ppf}(0.975, df=n-2) \cdot SE(\hat{\beta}_1) & \text{if } n \text{ is small} \end{cases}$$

Last time

We have n samples $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ generated from the following simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

The least-squares estimates of β_0 and β_1 are given by

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x} \text{ and } \widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

We can show

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

and hence

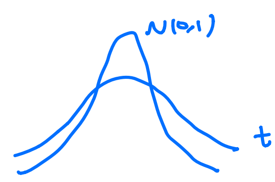
standardize

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

Since σ is an unknown parameter, we approximate it with the SD of the residuals, denoted as $\widehat{\sigma}$. A resulting statistic is $T = \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{SE}(\widehat{\beta}_1)}$ where

random

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



and a known fact is that

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

When n is large, the t-distribution with degree of freedom $n-2$ is close to the standard normal distribution, so

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

n is large

We can use the distribution of T to construct 95% CI for β_1 or conduct hypothesis testing $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$.

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)

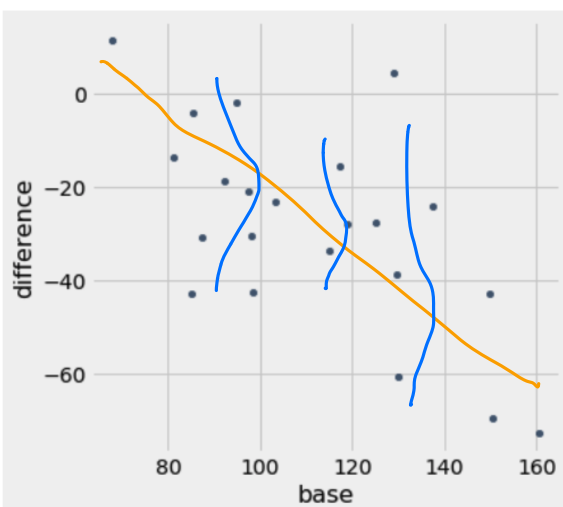
height	rad	chemo	base	month15	difference
164	679	180	160.57	87.77	-72.8
168	311	180	98.24	67.62	-30.62
173	388	239	129.04	133.33	4.29
157	370	168	85.41	81.28	-4.13
160	468	151	67.94	79.26	11.32
170	341	96	150.51	80.97	-69.54
163	453	134	129.88	69.24	-60.64
175	529	264	87.45	56.48	-30.97
185	392	240	149.84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

x Y

... (12 rows omitted)

$n=22$

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



$y=\beta_0+\beta_1 x$

Assumption for linear regression model?

$Y_i = \beta_0+\beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0,\sigma^2)$

OLS Regression Results

measures goodness of fit of linear regression model

Dep. Variable:	difference	R-squared:	0.397

$\parallel$ response

$R^2 \approx 1$: good fit

$R^2 \approx 0$: bad fit

	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 $\underbrace{\text{base health}}_{x}$ 100?

$\hat{\beta_0} = 32.1721, \ \hat{\beta_1} = -0.5447$

$\Rightarrow \ \hat{Y} = \hat{\beta_0} + \hat{\beta_1} x = \hat{\beta_0} + \hat{\beta_1} \cdot 100 = -22.3$

Multiple Regression

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

```
h_data.corr()
```

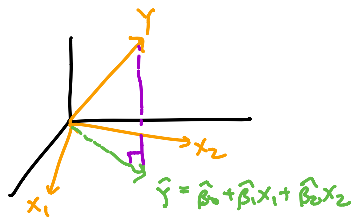
	height	rad	chemo	base	month1
height	1.000000	-0.305206	0.576825	0.354229	0.39052
rad	-0.305206	1.000000	-0.003739	0.096432	0.04061
chemo	0.576825	-0.003739	1.000000	0.062187	0.44578
base	0.354229	0.096432	0.062187	1.000000	0.56137
month15	0.390527	0.040616	0.445788	0.561371	1.00000
difference	-0.043394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

Conceptual picture:

Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ $\nearrow$ $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

Estimate $\beta_0, \beta_1, \beta_2$ by minimizing $\frac{1}{n}\sum_{i=1}^{n} (Y_i - a - b x_{1i} - c x_{2i})^2$ w.r.t. a, b, c

write $\underset{\text{vec}}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\underset{\text{vec}}{X_1} = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1n} \end{pmatrix}$, $\underset{\text{vec}}{X_2} = \begin{pmatrix} x_{21} \\ \vdots \\ x_{2n} \end{pmatrix}$:



OLS Regression Results

Dep. Variable:	difference	R-squared:	0.546

p-value

		coef	std err	t	P>|t|	[0.025	0.975]
β_0	const	-0.9992	20.227	-0.049	0.961	-43.335	41.336
β_1	base	-0.5655	0.134	-4.226	0.000	-0.846	-0.285
β_2	chemo	0.1898	0.076	2.500	0.022	0.031	0.349

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

R^2 is higher so better fit.

Conclude β_0 is not significantly different from 0

" β_1, β_2 are significantly different from 0

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

Chemo → Y
Height ↗

Height is correlated w/ {chemo, base}

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

a very minor improvement

OLS Regression Results

Dep. Variable:	difference	R-squared:	0.550

	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

Overview of Class

Ch 6: Measuring Variability

- You learned how the variance and SD is the average spread of your data from the mean. $= E((X-\mu_X)^2)$ $= E(X^2) - \mu_X^2$ $= \sqrt{Var(X)}$
- You should be able to compute Var(X) and SD(X) given a distribution table / a density function.
- If there are two random variables X and Y, and Y is a linear function of X, i.e. $Y = aX + b$, how to compute Var(Y) and SD(Y) from Var(X) and SD(X)? $= a^2 Var(X)$ $= |a| \cdot SD(X)$
- If we don't assume anything about the population distribution except the mean and SD, you can use Chebyshev's inequality to get an upper bound on the tail probability.

Ch 7: The Variance of a Sum

- If two random variables X and Y are independent, Var($X + Y$) is given by the sum of Var(X) and Var(Y).
- If $X_1, X_2, \ldots, X_n$ are i.i.d. samples from a population distribution and $S = X_1 + \cdots + X_n$ is the sample sum, Var(S) $= n\sigma^2$ and SD(S) $= \sqrt{n}\sigma$, where $\sigma = \text{SD}(X)$.
- The law of large number says the sample mean $\bar{X}$ converges to $\mu = E(X)$ as the sample size n grows. In particular, we proved the weak law of large numbers using Chebyshev's inequality. $P(|\bar{X}-\mu| \geq \varepsilon) \to 0$ as $n \to \infty$

Ch 8: The Central Limit Theorem

- The central limit theorem (CLT) says the distribution of sample mean $\bar{X}$ is "always" approximately normal, i.e. $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ when n is large enough.
  regardless of the population distr/n
- For any random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we can define a new random variable $X^\star$, called X in standard units, as $X^\star = \frac{X-\mu}{\sigma}$. $X^\star$ follows a standard normal distribution $\mathcal{N}(0, 1)$. The CDF of the standard normal distribution is written as $\Phi(x) = P(X^\star \leq x)$.

Ch 9: Inference

- Given n i.i.d. samples from a population distribution (say Bernoulli, normal, etc), you learned how to estimate the population parameter such as the population mean or population proportion.
- From our samples, we can make hypotheses about the value of the parameter of the population distribution. Assuming the null hypothesis is true, we compute a test statistic and compute the p-value. If the p-value is less that 0.05 at level 5%, we reject the null.
- A 95% CI for an unknown parameter tells you the rough uncertainty of the parameter.
- You should be able to conduct hypothesis testing for the population mean (both for one-sided and two-sided alternative hypotheses) and also construct 95% confidence interval.
- "Interpretation of CI" is important and you should be able to tell what is the right/wrong interpretation.

Ch 10: Probability Density

- For a continuous random variable, we compute the probability, expectation, and variance using the probability density function.

  Properties of density, CDF $= F(x) = P(X \le x)$ $\longleftrightarrow$ density $f(x) = \frac{dF(x)}{dx}$

- The exponential distribution is used to model the random life time of an object.

  Expectation, half-life, memoryless property, CDF of $\max\{X_1, \dots, X_n\}$, $\min\{X_1, \dots, X_n\}$

- You should be able to perform hypothesis testing and construct confidence interval for the difference between two groups.

  A μ_A $\mu_A - \mu_B$

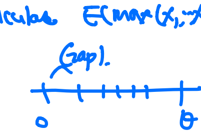
  B μ_B

Ch 11: Bias, Variance, and Least Squares

- The mean squared error (MSE) can be decomposed into the squared bias + variance.

  $MSE(T) = Bias^2(T) + Var(T)$

- The German Tank Problem discusses the parameter estimation for the discrete uniform distribution case, Unif$\{1, 2, \dots, N\}$. In class, we also discussed a similar problem for the continuous uniform distribution case, Unif$(0, \theta)$. Important to understand how to calculate $E(\max(X_1, \dots, X_n))$



- In regression, we aim to predict Y from a linear function of X, i.e. $\widehat{Y} = \widehat{a}X + \widehat{b}$. It is important to understand the properties of correlation and the residuals and its connection to regression.

  $r = r(X, Y)$ $D = Y - \widehat{Y}$

Ch 12: Inference in Regression

- The simple linear regression model assumes $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Looking at the scatter plot, can you determine whether the linear regression model is satisfied?
- You learned how to use the regression line $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ to estimate the true linear curve $\beta_0 + \beta_1 x_i$.
- You should be able to compute statistic/quantities and conduct hypothesis testing from the python output of the linear regression.

@ lec 39/40.