Week1

Prob

- Every event, Pr(A) >= 0
- Some event, Pr(total) = 1
- If A and B mutually exclusive, P(A or B) = P(A) + P(B)
- P(A) + P(Ac) = 1

Random Variables and Distributions

- Discrete
- Continuous
- PDF prob function for every random variables
- CDF F(x) = P(X <= x)
  - Pr(y1 <= Y <= y2) = F(y2) - F(y1)
  - **Non-decreasing!**

Mean(E(x))

- weighted average (sum(prob(x) * x))
- **Unique**
- *most frequently used measure of centrality*

Mode

- Value occuring with the greatest prob

Median

- 50% Median 50%
- Mislead under certain settings(skewed distribution)

Week 2

Variance

- The *variance* of a random variable measures the spread or dispersion of the variable around its mean

$$Var(X) = E\left[(X - \mu_X)^2\right]$$

- In the discrete case, this is simply the weighted average of the squared deviations of X from its mean

$$Var(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 \Pr(x_i)$$

Standard Deviation(SD)

- square root of Var (**X**)

Joint Distribution

- P(X = x, Y = y) = f(X,Y)

Marginal Distribution

- P(Y = y) = SUM(P(X = xi, Y = y)) for every xi

Conditional Distribution

P(Y = y | X = x) = P(x,y)/P(x)

Conditional Expectation

E[Y|X = x] = SUM(yi * prob(P(Y = yi|X = x))) for every yi

Independence

- X and Y are said to be independent if P(Y = y| X =x) = P(Y= y) = P(X = x)
- If x, y are independent, then P(X = x, Y = y) = P(X = x)*P(Y =y)

Dependence and Covariance

Cov(X, Y) = E[(X-mu(x))(Y - mu(y))]

or the sum of any x and any y for (xi - mu(x))(yi - mu(y))P(X = xi, Y = yi)

Correlation

rou(X,Y) = cov(X,Y)/sd(X)sd(Y)

- Range : [-1,1]
- 0: no relation
- 1,-1: perfect pos/neg relation
- **If two variables are independent, they must then also be uncorrelated**
- ***However, A covariance / correlation of 0 does NOT imply independence***

Useful Distribution

Discrete : *Bernoulli*

Continuous: Uniform, Normal, Standard Normal

- Bernouli
  - P(x = 1) = p
  - P(X = 0) = 1-p
  - E(X) = p and Var(x) = p(1-p)
- Uniform
  - pdf: f(y) = 1/(b-a)
  - Cdf: F(y) = (y-a)/(b-a)
- Normal, standard Normal
  - Standard normal (Z = (Y-mu)/sigma)~N(0,1) -> mean = 0, sd = 1
  - Qnorm(prob), pnorm(z-value)

Week 3

Random sampling

- An ideal way of collecting and constructing the samples.

Point Estimate

- Use sample mean X_bar to estimate population mean Mu_x
- Use sample mean X_var to estimate population var var_x
- Use sample cover sXY to estimate population coverage covXY

Mean, Var and standard error of sample mean

1. Mean equal to the population mean.

$$E(\bar{X}) = \mu_X$$

2. Variance equal to the population variance divided by the sample size.

$$Var(\bar{X}) = \frac{\sigma_X^2}{n}$$

- Specifically, the standard error for a sample of size n is

$$SE\left[\bar{X}\right] = \frac{\sigma_X}{\sqrt{n}}$$

or (simplifying notation)

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Large Num law and CLT

LLN : X_bar -> mu_X as n-> inf

CLT: normalized z-value -> N(0,1)

Property of Expectation and Var, Sd

- E(aX) = aE(x)
- E(X +b) = E(x) + b
- E(X + Y) = E(X) + E(Y)
- Var(aX) = a^2Var(X)
- Var(X + b) = Var(X)

- sd(aX) = |a|*sd(X)
- sd(X + b) = sd(X)

If X,Y dependent

- For two variables, X and Y we have

$$Var(X+Y) = Var(X) + Var(Y) + 2 \cdot Cov(X,Y)$$

- For three variables, X, Y, and Z we have

$$Var(X+Y+Z) = Var(X) + Var(Y) + Var(Z) + 2 \cdot Cov(X,Y) + 2 \cdot Cov(X,Z) + 2 \cdot Cov(Y,Z)$$

Or Var(X) + Var(Y)

- Lastly, for **constants** a and b

$$\begin{aligned} Cov(X,Y) &= Cov(Y,X) \\ Cov(aX,bY) &= abCov(X,Y) \\ Corr(aX,bY) &= Corr(X,Y) \end{aligned}$$

- For example,

$$\begin{aligned} Var(aX+bY) &= Var(aX) + Var(bX) + 2Cov(aX,bY) \\ &= a^2 Var(X) + b^2 Var(Y) + 2abCov(X,Y) \end{aligned}$$

Week4

- CI vs PI:
  - CI is the pred of sample mean, which is influenced by the sample size
  - PI is the pred of each value, not affected by size
  - **Pay attention to the one-side / two-side!**
- Null Hypothese
  - Assume H0 and HA
  - calculate what possiblity we can reject the hypo
    - H0, HA
    - Use sample mean, H0, sd of sample and size of sample
      - t = (sample mean-H0)/(sd/sqrt(size))
      - One-tail/two-tail, compare phi(t) with significant level.

- Comparing means for different populations
    -
        - Let's consider the two-sided null hypothesis than men and women are paid the same:

$$H_0 : \mu_m - \mu_w = 0 \text{ vs. } H_A : \mu_m - \mu_w \neq 0$$

- To test this, we first need an estimate of the difference in population means: $\mu_m - \mu_w$
- Of course, this is simply the difference in sample means: $\overline{Y}_m - \overline{Y}_w$
- We also need to know the standard error of this estimate.
- It can be computed as

$$SE(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

Aside: can you see why this is? Hint: what is $var(A - B)$ if A and B are independent?

- The t-statistic for testing H_0 is constructed the same way as before

$$t^{act} = \frac{(\overline{Y}_m - \overline{Y}_w) - 0}{SE(\overline{Y}_m - \overline{Y}_w)}$$

only now we call it "the t-statistic for comparing two means".

- Once again, using the CLT, it can be proven that this t-statistic is asymptotically standard normal under the null.
- Therefore, the p-value for the two-sided test can be computed exactly as before:

$$p\text{-value} = 2 \cdot \Phi\left(-\left|t^{act}\right|\right)$$

- Alternatively, a 95% CI for the difference in means can be computed as

$$(\overline{Y}_m - \overline{Y}_w) \pm 1.96 \cdot SE(\overline{Y}_m - \overline{Y}_w)$$

    -

Week5

OLS: predict b1 and b0:

- Sxy is the **sample covariance**
- Sx2 is the **sample variance** of x

$$\widehat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

Standard error of the slope:

SE(B1) = SER/(sqrt(n)*sqrt(Var(x))

SE(B0) = SER*sqrt(1/n + mean(x)^2/(n*Var(x)))

- The standard error R reports for $\widehat{\beta}_1$ is computed as[1]

$$SE(\widehat{\beta}_1) = \frac{SER}{\sqrt{\sum (X_i - \overline{X})^2}} = \frac{SER}{\sqrt{n}\sqrt{\frac{\sum (X_i - \overline{X})^2}{n}}}$$

- The standard error R reports for $\widehat{\beta}_0$ is computed as

$$SE(\widehat{\beta}_0) = \frac{SER}{\sqrt{\sum (X_i - \overline{X})^2}} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2} = SER \cdot \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2}}$$

Week6

Goodness of fit:

SSR: **Sum of squared residuals**

SER: **Standard error of the regression**

- Instead, the *Standard Error of the Regression* (SER) is an estimator[2] of the standard deviation of e_i.

$$SER = s_{\widehat{e}} = \sqrt{\tfrac{1}{n-2}\sum \widehat{e}_i^2} = \sqrt{\tfrac{1}{n-2}\sum \left(Y_i - \widehat{Y}_i\right)^2} = \sqrt{\tfrac{SSR}{n-2}}$$

TSS: **Total sum of Squares**

- We can normalize the SSR using a measure of the *total* variation in Y, called the *Total Sum of Squares*:

$$TSS = \sum \left(Y_i - \overline{Y}\right)^2$$

R2 = 1 - SSR/TSS

$$R^2 = \frac{\Sigma\left(\widehat{Y}_i - \overline{Y}\right)^2}{\Sigma\left(Y_i - \overline{Y}\right)^2} = \frac{ESS}{TSS} = \frac{\text{“explained variation”}}{\text{“total variation”}}$$

R2 = rxy2

- **IMPORTANT : R2(x,y) = R2(5x,10y)**

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

Attention

- *A high R**2 means that a lot of the total variation is explained by the regression*

  (data is tightly concentrated around the line).
- *R**2 is a measure of fit of the **linear** model, so it can miss non-linear relationships*
- *R**2 **does not prove** that our model is right or wrong:*

  *You can have a good model but a low R**2 because Var (**e**i) is large.*
-
  - You can also have a bad model with $R^2 \approx 1$.
    - Spurious correlation/regression: X & Y move together because of something else. (remember ice cream and muder?!?!)
  - Finally, R^2 does not tell you the direction: it could also be $Y \rightarrow X$ (reserve causation).

F statistics

- The F-statistic is constructed as follows:

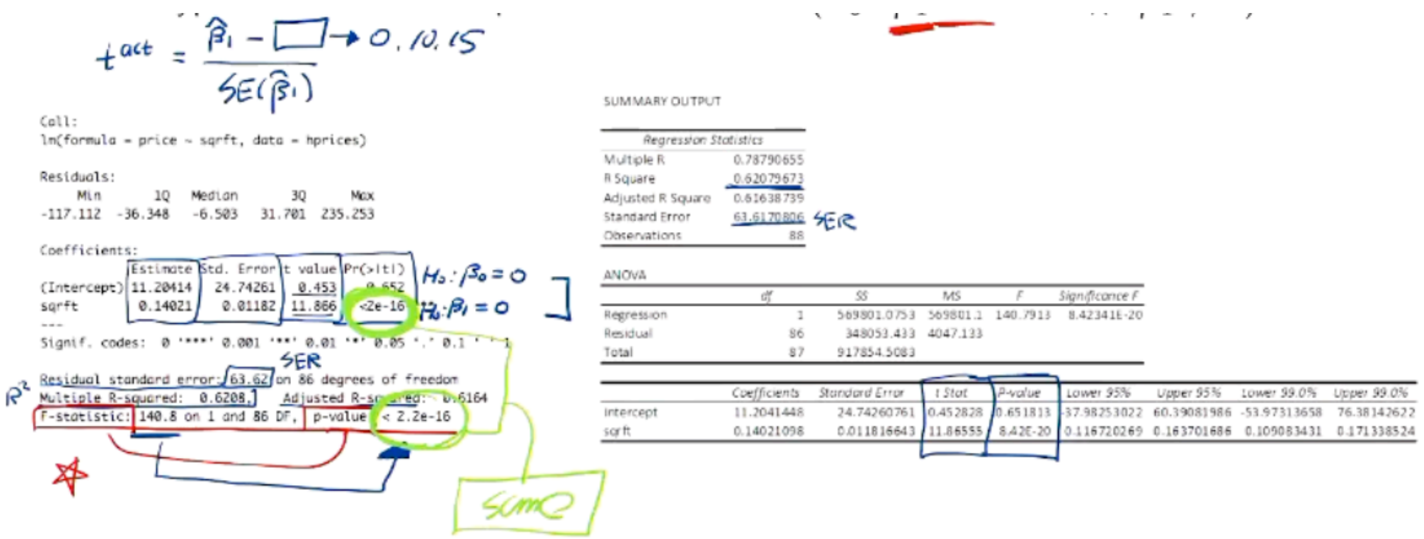
$$F = \frac{(n-2)\,R^2}{(1-R^2)}$$

- **If** $e_i \sim N\left(0, \sigma^2\right)$ it can be shown that

$$F \xrightarrow{d} F_{1,n-2}$$

so you can calculate p-value using $1 - pf\left(F, 1, n-2\right)$ command in R.

- In the **univariate** regression case, F-statistic conveys **exactly**[7] the same information as the t-stat for β_1 (t-stat for the slope parameter coefficient). That is, $t^2_{\beta_1} = F$.



$$t^{act} = \frac{\hat{\beta}_1 - \boxed{} \to 0, 10, 15}{SE(\hat{\beta}_1)}$$

```
Call:
lm(formula = price ~ sqrft, data = hprices)

Residuals:
     Min       1Q   Median       3Q      Max
-117.112  -36.348   -6.503   31.701  235.253

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.20414   24.74261   0.453    0.652
sqrft        0.14021    0.01182  11.866   <2e-16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.62 on 86 degrees of freedom
Multiple R-squared:  0.6208,    Adjusted R-squared:  0.6164
F-statistic: 140.8 on 1 and 86 DF,  p-value: < 2.2e-16
```

$H_0: \beta_0 = 0$

$H_0: \beta_1 = 0$

SER

R^2

same

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.78790655
R Square	0.62079673
Adjusted R Square	0.61638739
Standard Error	63.6170806
Observations	88

SER

ANOVA

	df	SS	MS	F	Significance F
Regression	1	569801.0753	569801.1	140.7913	8.42341E-20
Residual	86	348053.433	4047.133		
Total	87	917854.5083			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	11.2041448	24.74260761	0.452828	0.651813	-37.98253022	60.39081986	-53.97313658	76.38142622
sqrft	0.14021098	0.011816643	11.86555	8.42E-20	0.116720269	0.163701686	0.109083431	0.171338524

Dummy variable(binary x variable(0,1))

- B1: doesn't mean slope here, but the difference between two groups(0,1)
- B0 = Intercept = f(0)
- f(1) = Intercept+B1

- Here's the output from the regression using "Male" as a binary variable.

```
> fit = lm(earnings~male, data=cps12)
> summary(fit)

Call:
lm(formula = earnings ~ male, data = cps12)

Residuals:
    Min      1Q  Median      3Q     Max
-23.162  -7.560  -2.222   5.421  51.624

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.5024     0.2514   85.54   <2e-16 ***
male          3.7965     0.3531   10.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.1 on 3953 degrees of freedom
Multiple R-squared:  0.02841,   Adjusted R-squared:  0.02816
F-statistic: 115.6 on 1 and 3953 DF,  p-value: < 2.2e-16

> confint.default(fit, level=0.95)
               2.5 %   97.5 %
(Intercept) 21.009706 21.99505
male         3.104362  4.48860
```

- Dataset: cps12.csv
- $\widehat{\beta}_0 = 21.50$ is the sample average of earnings for women
- $\widehat{\beta}_0 + \widehat{\beta}_1 = 25.30$ is the sample average of earnings for men.
- $\widehat{\beta}_1 = 3.80$ is the difference between these two sample averages.

-

OLS Assumptions(*ordinary least squares*)

- Do the OLS formulas have the same desirable properties that $\overline{X}$ had?
  1. Unbiasedness: $E(\widehat{\beta}_i) = \beta_i$
  2. Consistency: $\widehat{\beta}_i \xrightarrow{p} \beta_i$
  3. Asymptotic normality: $\widehat{\beta}_i$ distributed Normally for large n

OLS Assumption 1 **Linearity in parameters; zero conditional mean**

The true model is $Y_i = \beta_0 + \beta_1 X_i + u_i$ **and** $E(u_i \mid X_i) = 0$.

OLS Assumption 2 **Simple random sample**

(X_i, Y_i) are *iid* draws from their joint distribution.

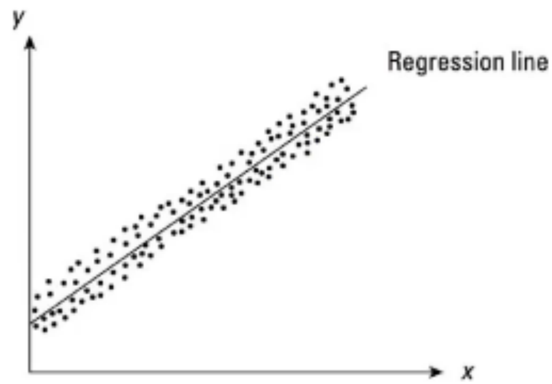
OLS Assumption 3 **No extreme outliers**

u_i and X_i have non-zero & finite fourth moments.

$$0 < E\left(X_i^4\right) < \infty \text{ and } 0 < E\left(u_i^4\right) < \infty$$

A4

Homoskedastic: for all values of X, the var of error is the same

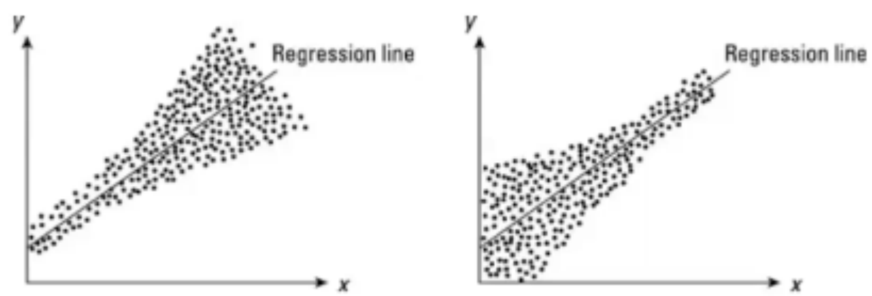


- If $Var(u_i \mid X_i) = \sigma_u^2$ (a constant), then we have homoskedasticity.
  - This is OLS Assumption 4.

- If instead $Var(u_i \mid X_i) = f(X_i)$, we have heteroskedasticity.
  - This violates OLS Assumption 4

Heteroskedasticity:

- However, in practice we might see many cases like this:



- If errors are Peter, B0 and B1 are correct
- But the SE formulas changes to a more complicated one

```
cps12 = read.csv(url("http://hanachoi.github.io/datasets/cps12.csv"), header=TRUE,
sep=",") # load cps dataset
fit_cps12 = lm(earnings~male, data=cps12) # run regression of earnings on male dummy
variable
fit_cps12$HRse = vcovHC(fit_cps12, type="HC1") # obtain HR SEs
coeftest(fit_cps12) # report homoskedastic SE
coeftest(fit_cps12, fit_cps12$HRse) # report HR SE
```

- We can use heter one to do data analysis

Multiple Regression

In particular, if the true model is

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + u$$

but you leave out X_2 and estimate

$$Y = \beta_0 + \beta_1 X_1 + u$$

then you can expect the following biases[3] in your estimate of β_1.

	$\rho_{X_1X_2} > 0$	$\rho_{X_1X_2} < 0$
$\alpha_2 > 0$	positive bias	negative bias
$\alpha_2 < 0$	negative bias	positive bias

The OLS Assumptions in the Multiple Regression Model

OLS Assumption 1 Linearity in parameters; zero conditional mean

$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + u_i$ & $E(u_i \mid X_{1i}, \ldots, X_{pi}) = 0$

OLS Assumption 2 Simple random sample

$(Y_i, X_{1i}, \ldots, X_{pi}) \sim iid$

OLS Assumption 3 No extreme outliers

$X_{1i}, \ldots, X_{pi}, u_i$ have nonzero, finite fourth moments.

OLS Assumption 4 No perfect collinearity

Regressors can't be written as linear combinations of each other.

OLS Assumption 5 Homoskedasticity

$Var(u_i \mid X_{1i}, \ldots, X_{pi}) = \sigma_u^2$

F-statistics

$$F = \frac{(n-p-1)\, R^2}{p\,(1-R^2)}$$

Goodness of Fit

- SER
- R^2
- Adj-R2

$$SER = s_{\hat{u}} = \sqrt{\frac{SSR}{n-p-1}} = \sqrt{MSE}$$

$$R^2 = \frac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2}$$

$$\overline{R}^2 = 1 - \frac{n-1}{n-p-1}\frac{SSR}{TSS}$$

- adj R square can be negative
- if adj R square is far apart from R2, it's a bad sign.
-