Chapter 5 Confidence Interval, t Distribution

- X bar and X has different Z standardization because X bar use standard error (divided by square root of n) but not for X.

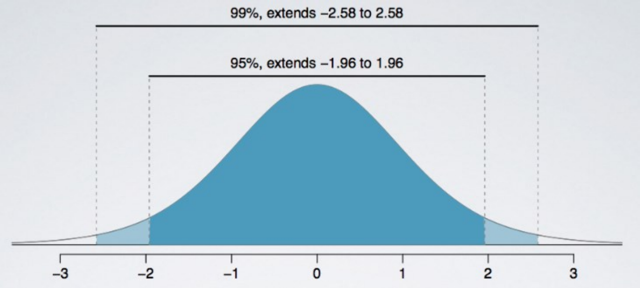
$$
\begin{aligned}
P(6 \le \bar{X} \le 8) &= P\left(\frac{6-7}{\sqrt{\frac{25}{100}}} \le Z \le \frac{8-7}{\sqrt{\frac{25}{100}}}\right)\\
&= P\left(-\frac{1}{5/10} \le Z \le \frac{1}{5/10}\right)\\
&= P(-2 \le Z \le 2)\\
&= 0.9772-(1-0.9772)\\
&= 0.9544
\end{aligned}
$$

$$
\begin{aligned}
P(6 \le X \le 8) &= P\left(\frac{6-7}{\sqrt{25}} \le Z \le \frac{8-7}{\sqrt{25}}\right)\\
&= P\left(-\frac{1}{5} \le Z \le \frac{1}{5}\right)\\
&= 0.5793-(1-0.5793)\\
&= 0.1586
\end{aligned}
$$

- Predictive interval calculated using the SD and Confidence interval calculated using the Standard error. Z value is x bar – miu divided by standard error of x bar. To calculate z-score and compare with 95 confidence interval or using mean +- 2se.

$$z = (x - \mu) / (\sigma / \sqrt{n})$$

- Why there is question in the additional exercise that predictive interval that calculated using standard error?
- The null hypothesis rejected if the z value is out of range -1.96 < z < 1.96.
- Statistical Inference: The i.i.d. sampling each of the following is true except E(Y) < E(Y)
- Confidence Interval (Q5 2013)



$$P(0 < Z < 1) = .34$$

$$P(-1 < Z < 1) = .68$$

$$P(-2 < Z < 2) = .954$$

$$P(-1.96 < Z < 1.96) = .95$$

$$P(-3 < Z < 3) = .9974$$

- Z critical values

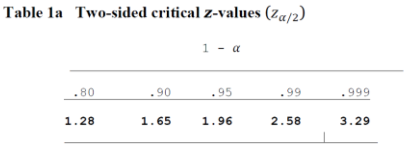


Table 1a Two-sided critical z-values ($z_{\alpha/2}$)

$1-\alpha$				
.80	.90	.95	.99	.999
1.28	1.65	1.96	2.58	3.29

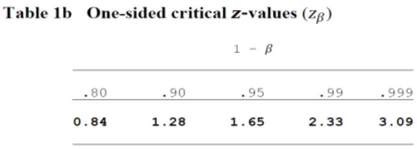
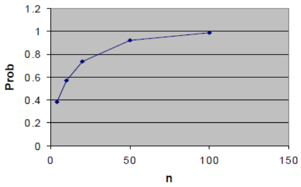


Table 1b One-sided critical z-values (z_{β})

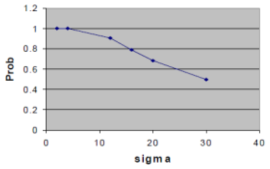
$1-\beta$				
.80	.90	.95	.99	.999
0.84	1.28	1.65	2.33	3.09

- 95% Confidence Interval of Coefficient Variables (Beta) is = Beta – 1.96 x se, Beta + 1.96 x se.
- The probabilities calculated from the interval, getting higher if n is higher. On the other hand, the sigma larger, probability lower.



$$f(n) = P\left(195 \le \bar{X} \le 205\right) = P\left(\frac{195-200}{\frac{20}{\sqrt{n}}} \le Z \le \frac{205-200}{\frac{20}{\sqrt{n}}}\right)$$

$$f(n) = P\left(-\frac{\sqrt{n}}{4} \le Z \le \frac{\sqrt{n}}{4}\right)$$



- If you want to see the result, just plug in any number in the model and compare 2 results. However, there are many other factors that affect dependent variables. Note that predictive interval is really big (look at the estimate of the variance of residuals (3865666562.7)/207.
- Unit of measurement influence beta, let say convert 1 to 1000 UoM, then Beta x 1000.
- **Interpret the p-values.** People seemed very confused at the tutorial in respect to p-values. You only have to compare them to significance levels:

90% confidence	95% confidence	99% confidence
α = .10	α = .05	α = .01

So if you are testing at a 90% confidence, you have a significance level α = .10. To test the null hypothesis at level α, we reject if the p-value is less than α. So in our example, p-value = .09 therefore I do not reject at 5% nor 1%. However .09 < .10 so I marginally reject the null at 10%. p-value that is zero to at least four decimal places.

- In order to know how many samples needed for the research, we need to calculate this way

$$We\,want\; P\left(-5 \le X - \mu \le 5\right) = 0.99, given\, \sigma_X = 15$$

$$In\,general\; P\left[-Z\left(1-\frac{\alpha}{2}\right)\sigma_{\bar{X}} \le \bar{X} - \mu \le Z\left(1-\frac{\alpha}{2}\right)\sigma_{\bar{X}}\right] = 1-\alpha$$

$$Thus, 1-\alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \frac{\alpha}{2} = 0.005 \Rightarrow 1-\frac{\alpha}{2} = 0.995$$

$$So, P\left(-Z_{.995}\frac{15}{\sqrt{n}} \le \bar{X} - \mu \le Z_{.995}\frac{15}{\sqrt{n}}\right) = 0.99$$

$$\Rightarrow Z_{.995}\frac{15}{\sqrt{n}} = 5$$

$$\Rightarrow \quad n = \left(Z_{.995}\frac{15}{5}\right)^2$$

$$= \left(2.576 \cdot \frac{15}{5}\right)^2$$

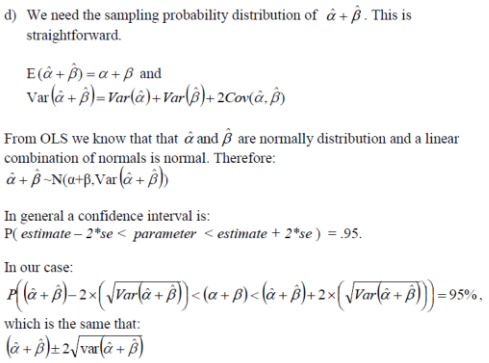
$$= 59.7\ (round\ up) \approx 60\, observations$$

- Read Stats output

```
      Source |       SS       df       MS              Number of obs =     209
-------------+------------------------------           F(  1,   207) =    2.77
       Model |  5166419.33     1  5166419.33           Prob > F      =  0.0978
    Residual |   386566563   207  1867471.32           R-squared     =  0.0132
-------------+------------------------------           Adj R-squared =  0.0084
       Total |   391732982   208  1883331.64           Root MSE      =  1366.6

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |   18.50119   11.12325     1.66   0.098    -3.428195    40.43057
       _cons |   963.1913   213.2403     4.52   0.000     542.7902    1383.592
------------------------------------------------------------------------------
```

α: If the return on equity is zero, roe=0, then the predicted salary is the intercept, 963.191 which equals $963,191 since salary is measured in thousands. β: If the return on equity increases by one percentage point, then the salary is predicted to change by about 18.50 or $18,501.

- Important logic for the interval



d) We need the sampling probability distribution of $\hat{\alpha}+\hat{\beta}$. This is straightforward.

$\mathrm{E}(\hat{\alpha}+\hat{\beta})=\alpha+\beta$ and
$\mathrm{Var}(\hat{\alpha}+\hat{\beta})=Var(\hat{\alpha})+Var(\hat{\beta})+2Cov(\hat{\alpha},\hat{\beta})$

From OLS we know that that $\hat{\alpha}$ and $\hat{\beta}$ are normally distribution and a linear combination of normals is normal. Therefore:
$\hat{\alpha}+\hat{\beta} \sim \mathrm{N}(\alpha+\beta, \mathrm{Var}(\hat{\alpha}+\hat{\beta}))$

In general a confidence interval is:
P(*estimate – 2*se* < *parameter* < *estimate + 2*se*) = .95.

In our case:
$$P\left((\hat{\alpha}+\hat{\beta})-2\times\left(\sqrt{Var(\hat{\alpha}+\hat{\beta})}\right)<(\alpha+\beta)<(\hat{\alpha}+\hat{\beta})+2\times\left(\sqrt{Var(\hat{\alpha}+\hat{\beta})}\right)\right)=95\%,$$
which is the same that:
$$(\hat{\alpha}+\hat{\beta})\pm 2\sqrt{\mathrm{var}(\hat{\alpha}+\hat{\beta})}$$

- Two sided or one sided is matter. More or less is one sided, In between is two sided, does not equal is two sided (absolute value)

Use **qt(1 - α/2, df)** for 2-sided critical *t*-value

- Look for this table to calculate t value if n is small. Get information of n and also confidence interval.

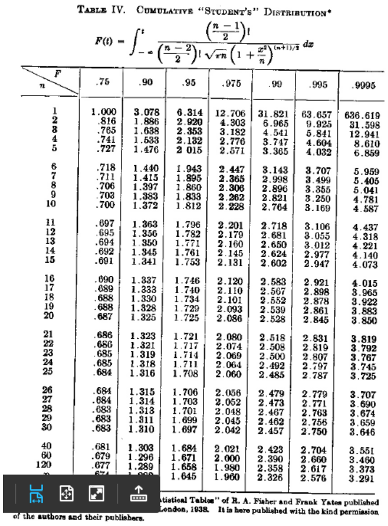


TABLE IV. CUMULATIVE "STUDENT'S" DISTRIBUTION*

$$F(t)=\int_{-\infty}^{t}\frac{\left(\frac{n-1}{2}\right)!}{\left(\frac{n-2}{2}\right)!\sqrt{\pi n}\left(1+\frac{x^2}{n}\right)^{(n+1)/2}}\,dx$$

n \ F	.75	.90	.95	.975	.99	.995	.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.886	2.920	4.303	6.965	9.925	31.598
3	.765	1.638	2.353	3.182	4.541	5.841	12.941
4	.741	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	1.476	2 015	2.571	3.365	4.032	6.859
6	.718	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	1.415	1.895	2.365	2.998	3.499	5.405
8	.706	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	1.319	1.714	2.069	2.500	2.807	3.767
24	.685	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	1.303	1.684	2.021	2.423	2.704	3.551
60	.679	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	1.289	1.658	1.980	2.358	2.617	3.373
[illegible]	[illegible]	[illegible]	1.645	1.960	2.326	2.576	3.291

[illegible] atistical Tables" of R. A. Fisher and Frank Yates published [illegible] London, 1938. It is here published with the kind permission of the authors and their publishers.

Chapter 6 Ordinary Least Square

- Obtain beta from the manual calculation is slope = covariance XY / variance X
- The OLS estimator is biased if the omitted variable is correlated with the included variable. Because then the omitted variable will be absorbed by error.

ASSUMPTION	DISTRIBUTION
n large and σ_X^2 known	$\frac{\overline{X}-\mu_X}{\sigma_X/\sqrt{n}} \sim N(0,1)$
$X_i \sim N(\mu_x, \sigma_X^2)$ and σ_X^2 known	$\frac{\overline{X}-\mu_X}{\sigma_X/\sqrt{n}} \sim N(0,1)$
n large and σ_X^2 unknown	$\frac{\overline{X}-\mu_X}{s_X/\sqrt{n}} \sim N(0,1)$
n small, $X_i \sim N(\mu_x, \sigma_X^2)$ and σ_X^2 unknown	$\frac{\overline{X}-\mu_X}{s_X/\sqrt{n}} \sim t_{n-1}$

- Degree of freedom, k calculated from the number of available variables (Beta) in the model.

$$T = \frac{Z}{\sqrt{X/k}}$$

$$T \sim t_k$$

"*T* is *t*-distributed with *k* degrees of freedom"

Chapter 7 Simple Linear Regression, Plugin Prediction

- Fitted Value vs Error

$$\varepsilon_i = y_i - (\alpha + \beta x_i) \approx y_i - (\hat{\alpha} + \hat{\beta} x_i) = e_i$$

$$\mathrm{e}_\mathrm{i} = y_i - \hat{y}_i$$

Chapter 8 Simple Linear Regression Fits Residual R Square

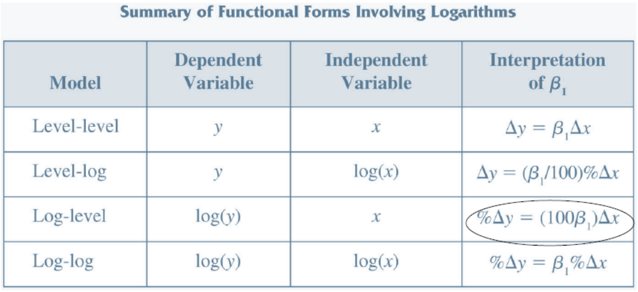
$$\boxed{s_y^2 = s_{\hat{y}}^2 + s_e^2 \quad \text{because resids and fits have 0 sample correlation.}}$$

$$\Rightarrow$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

total variation in y = variation explained by x + unexplained variation

SST = SSE + SSR



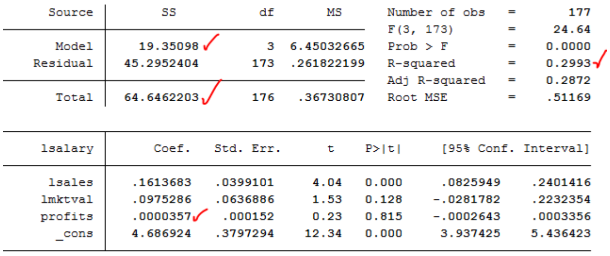
Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Chapter 9-10 Multiple Regression

- The error is independent (zero conditional mean assumption). We need error is independent. This is important for unbiasedness, to check causality or not. Umbrellas vs Rain. There is related but no causality. In finance, there is causality.
- Goodness of fit (R Square) should not depend on the units of measurement
- Understand the relationship between variables with negative or positive coefficient.
- Concern about a large ceteris paribus difference in independent variables – measures if it almost two and one-half standard deviations – is needed to obtain a predicted difference in dependent variables or a half a point.
- The variables having negative values cannot be converted to logarithm like profits.

- How to measure percentage increase in explanation of additional variables by percentages? by just looking at the coefficient from regression.
- Learning how to construct model
- The more efficient model can be tested by comparing variance number, lower variance is more efficient. A more efficient estimator has a smaller variance
- The sample correlation between log (mktval) and profits is about .78, which is fairly high. As we know, this causes no bias in the OLS estimators, although it can cause their variances to be large. Given the fairly substantial correlation between market value and firm profits, it is not too surprising that the latter adds nothing to explaining CEO salaries. Also, profits is a short term measure of how the firm is doing while mktval is based on past, current, and expected future profitability. Do multicollinearity issue exist?
- To see the variables explain or not, test using the coefficient that is small or big and how many percentages explaining dependant variables. Compare this coefficient how many change in particular variables can change the dependent variables. Look at the r square to compare with how many percentage all variables explaining dependent one.



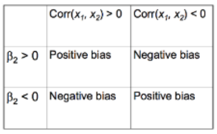
```
      Source |       SS       df       MS              Number of obs =     177
-------------+------------------------------           F(3, 173)     =   24.64
       Model |  19.35098       3  6.45032665           Prob > F      =  0.0000
    Residual |  45.2952404   173  .261822199           R-squared     =  0.2993
-------------+------------------------------           Adj R-squared =  0.2872
       Total |  64.6462203   176  .36730807            Root MSE      =  .51169

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lsales |   .1613683   .0399101     4.04   0.000     .0825949    .2401416
     lmktval |   .0975286   .0636886     1.53   0.128    -.0281782    .2232354
     profits |   .0000357    .000152     0.23   0.815    -.0002643    .0003356
       _cons |   4.686924   .3797294    12.34   0.000     3.937425    5.436423
------------------------------------------------------------------------------
```

- Formula:

$$F \equiv \frac{\left(R_{ur}^2 - R_r^2\right)/q}{\left(1-R_{ur}^2\right)/(n-k-1)}, \text{ where again}$$

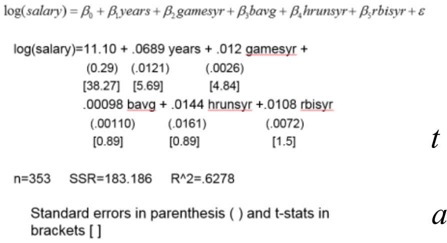
r is restricted and ur is unrestricted

$$\hat{\beta}_j : t - stat \equiv \hat{\beta}_j \Big/ se\left(\hat{\beta}_j\right)$$



	Corr(x_1, x_2) > 0	Corr(x_1, x_2) < 0
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

$$\text{H}_0\text{: } b_j = a_j$$



$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + \varepsilon$$

```
log(salary)=11.10 + .0689 years + .012 gamesyr +
           (0.29)   (.0121)        (.0026)
           [38.27]  [5.69]         [4.84]
           .00098 bavg + .0144 hrunsyr +.0108 rbisyr
            (.00110)     (.0161)        (.0072)
            [0.89]       [0.89]         [1.5]

n=353    SSR=183.186    R^2=.6278

  Standard errors in parenthesis ( ) and t-stats in
  brackets [ ]
```

$$t - stat = \left(\hat{\beta}_j - a_j\right) \Big/ se\left(\hat{\beta}_j\right), \text{ where}$$

$$a_j = 0 \text{ for the standard test}$$

$(1)\ H_a : \beta_j > 0$

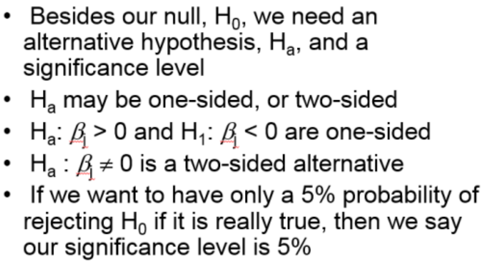
Reject H_0 if t - stat > critical value

$(2)\ H_a : \beta_j < 0$

Reject H_0 if t - stat < –(critical value)

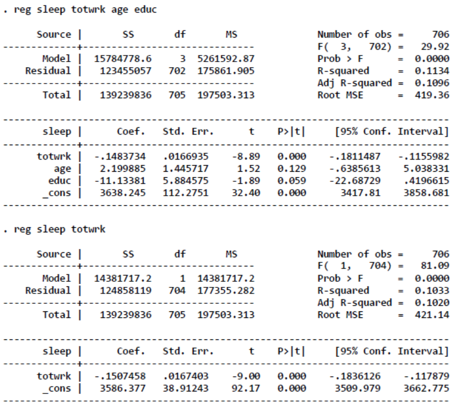
$(3)\ H_a : \beta_j \neq 0$

Reject H_0 if |t - stat| > critical value



- Besides our null, H_0, we need an alternative hypothesis, H_a, and a significance level
- H_a may be one-sided, or two-sided
- H_a: $\beta_1 > 0$ and H_1: $\beta_1 < 0$ are one-sided
- H_a : $\beta_1 \neq 0$ is a two-sided alternative
- If we want to have only a 5% probability of rejecting H_0 if it is really true, then we say our significance level is 5%

- Hypothesis Testing



```
. reg sleep totwrk age educ

      Source |       SS       df       MS              Number of obs =     706
-------------+------------------------------           F(  3,   702) =   29.92
       Model |  15784778.6     3  5261592.87           Prob > F      =  0.0000
    Residual |   123455057   702  175861.905           R-squared     =  0.1134
-------------+------------------------------           Adj R-squared =  0.1096
       Total |   139239836   705  197503.313           Root MSE      =  419.36

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1483734   .0166935    -8.89   0.000    -.1811487   -.1155982
         age |   2.199885   1.445717     1.52   0.129    -.6385613    5.038331
        educ |  -11.13381   5.884575    -1.89   0.059    -22.68729    .4196615
       _cons |   3638.245   112.2751    32.40   0.000      3417.81    3858.681
------------------------------------------------------------------------------

. reg sleep totwrk

      Source |       SS       df       MS              Number of obs =     706
-------------+------------------------------           F(  1,   704) =   81.09
       Model |  14381717.2     1  14381717.2           Prob > F      =  0.0000
    Residual |   124858119   704  177355.282           R-squared     =  0.1033
-------------+------------------------------           Adj R-squared =  0.1020
       Total |   139239836   705  197503.313           Root MSE      =  421.14

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1507458   .0167403    -9.00   0.000    -.1836126    -.117879
       _cons |   3586.377   38.91243    92.17   0.000     3509.979    3662.775
------------------------------------------------------------------------------
```

(e) With df = 706 – 4 = 702, we use the standard normal critical value, which is 1.96 for a two-tailed test at the 5% level. Now teduc = □11.13/5.88 □□□1.89, so |teduc| = 1.89 < 1.96, and we fail to reject H_0: $\beta_{educ} = 0$ at the 5% level. Also, $t_{age} \approx 1.52$, so age is also. Statistically insignificant at the 5% level. $t_{totwrk} = -.1483 \ / \ .0166 \approx -8.88$, because its absolute value is larger than 1.96, therefore totwrk is statistically significant. Failed to rejects means insignificant?

- If we get the R square low, we need to think other variables that might influence the dependent variables.
- F test formula, n is the number of observation, q is the number of tested joint variables (2) and k is the number of variables (3):

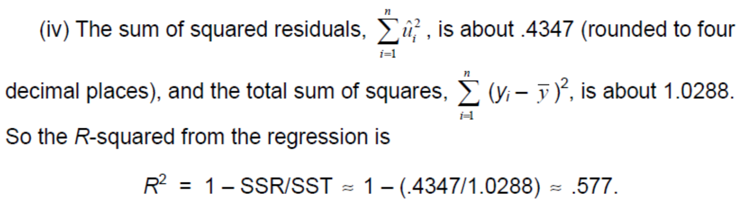
$$F \equiv \frac{\left(R_{ur}^2 - R_r^2\right)/q}{\left(1 - R_{ur}^2\right)/\left(n-k-1\right)}, \text{ where again}$$

r is restricted and ur is unrestricted

Since the R² from a model with only an intercept will be zero, the *F* statistic is simply

$$F = \frac{R^2/k}{\left(1-R^2\right)/\left(n-k-1\right)}$$

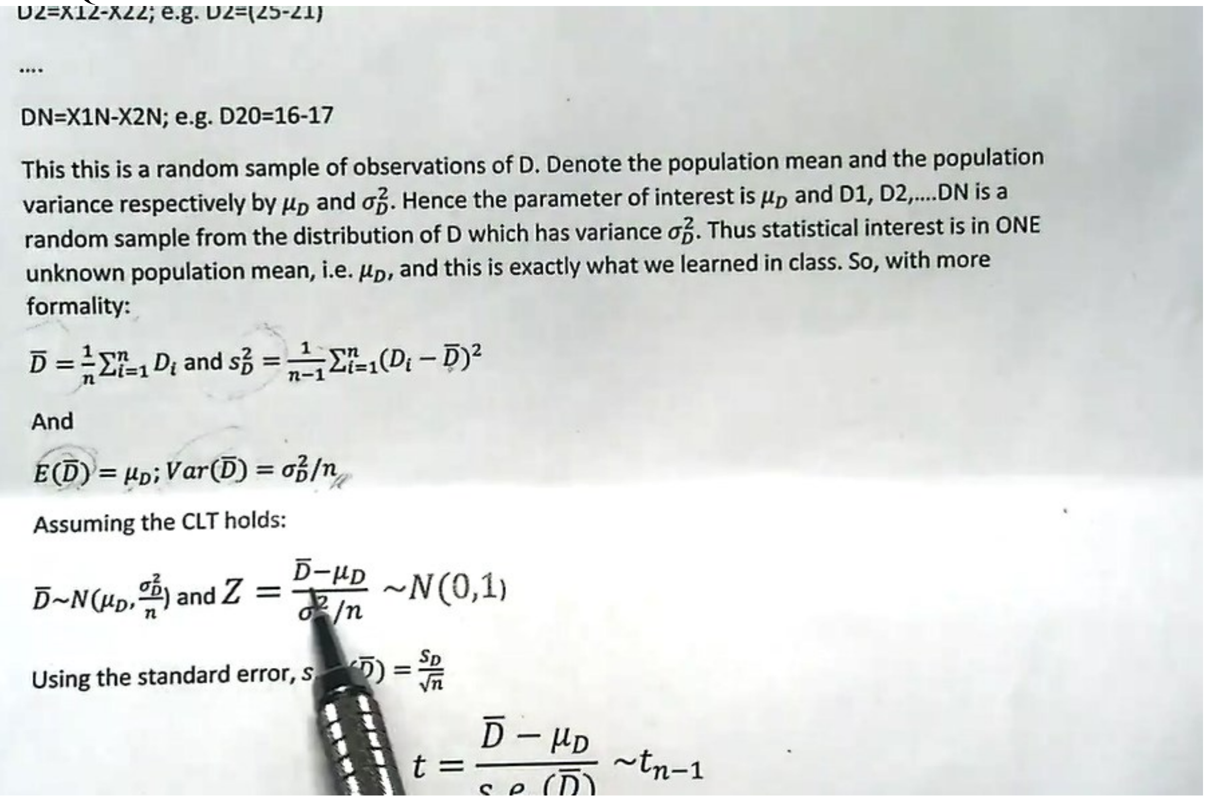
- We need to compute the R-squared form of the F statistic for joint significance. But F = [(.113 □□.103)/(1 □□.113)](702/2) □□3.96. The 5% critical value in the F(2,702) distribution with denominator df = □□has a critical value = 3.00. Therefore, educ and age are jointly significant at the 5% level (3.96 > 3.00). In fact, the p-value is about .019 (0.019 is the total p value of the joint variables), and so educ and age are jointly significant at the 2% level.
- Compare after and before omit variables for f test for the coefficients in the model.
- The standard t and F statistics that we used assume homoscedasticity, in addition to the other CLM assumptions. If there is heteroscedasticity in the equation, the tests are no longer valid. Heteroscedasticity tested using the scatter plot. Normal distribution tested using histogram.
- There is no reason to remove any of the explanatory variables from this model. These variables are individually significant, jointly significant (the p-value of the F-test is zero), high correlation between explanatory variables does not seem to be an issue, etc.
- All of the explanatory variables are statistically significant at the 5% level of significance (including the constant).
- This is how to calculate R square manually by not looking at Stata result.



(iv) The sum of squared residuals, $\sum_{i=1}^{n} \hat{u}_i^2$, is about .4347 (rounded to four decimal places), and the total sum of squares, $\sum_{i=1}^{n} (y_i - \bar{y})^2$, is about 1.0288. So the *R*-squared from the regression is

$$R^2 = 1 - \text{SSR/SST} \approx 1 - (.4347/1.0288) \approx .577.$$

- Exam Q



D2=X12-X22; e.g. D2=(25-21)

....

DN=X1N-X2N; e.g. D20=16-17

This this is a random sample of observations of D. Denote the population mean and the population variance respectively by μ_D and σ_D^2. Hence the parameter of interest is μ_D and D1, D2,....DN is a random sample from the distribution of D which has variance σ_D^2. Thus statistical interest is in ONE unknown population mean, i.e. μ_D, and this is exactly what we learned in class. So, with more formality:

$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ and $s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$

And

$E(\bar{D}) = \mu_D; Var(\bar{D}) = \sigma_D^2/n$

Assuming the CLT holds:

$\bar{D} \sim N(\mu_D, \frac{\sigma_D^2}{n})$ and $Z = \frac{\bar{D}-\mu_D}{\sigma_D^2/n} \sim N(0,1)$

Using the standard error, $s.e.(\bar{D}) = \frac{s_D}{\sqrt{n}}$

$$t = \frac{\bar{D} - \mu_D}{s.e.(\bar{D})} \sim t_{n-1}$$

- Only (ii), omitting an important variable, can cause bias, and this is true only when the omitted variable is correlated with the included explanatory variables. The homoskedasticity assumption, played no role in showing that the OLS estimators are unbiased. (Homoskedasticity was used to obtain the usual variance formulas for the ˆ*j* b.) Further, the degree of collinearity between the explanatory variables in the sample, even if it is reflected in a correlation as high as .95, does not affect the Gauss-Markov assumptions. Only if there is a *perfect* linear relationship among two or more explanatory the corresponding assumption is violated.
- Testing one variable by omitting another variable if the magnitude of simple linear regression for independent variables larger than the multiple regression.
- When to use one tailed when to use two tail?, Check again how to calculate standard error.
- The intercept unit of measurement follow the y head unit of measurement. Unit of slope is the unit of y head per unit of slope itself that is mentioned in the model.
- One tailed formula

$$P\left[-Z(1-\alpha)\sigma_{\bar{X}} + \bar{X} \le \mu\right] = 1-\alpha$$