Multiple Regression Analysis: Inference. Wooldridge (2013), Chapter 4 and Chapter 6 (section 6.4)

- Introduction
- Assumptions of the Classical Linear Model
- The t Test
- Hypothesis testing - one-sided alternatives
- Hypothesis testing - two-sided alternatives
- Confidence Intervals
- Testing a Linear Combination
- Multiple Linear Restrictions
- Testing Exclusion Restrictions
- The F statistic
- Overall Significance
- Prediction for the conditional mean of y
- Prediction for y
- Predicting y in a log model

Examples of test of hypothesis

Consider the model:

$$bwgth = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 npvis + \beta_4 age + u,$$

where

bwgth	-birth weight,
cigs	-cigarettes smoked per day while pregnant,
educ	-years of schooling for the mother,
npvis	-total number of prenatal visits,
age	-Age of the mother.

$$bwgth = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 npvis + \beta_4 age + u,$$

- Is the partial effect of *age* relevant after controlling for *cigs*, *education* and *npvis*?

$$H_0 : \beta_4 = 0 \ vs \ H_1 : \beta_4 \neq 0,$$

[Individual statistical significance]

- Is the effect of smoking 10 cigarettes canceled by the effect of one more prenatal visit?

$$H_0 : 10\beta_1 + \beta_3 = 0 \ vs \ H_1 : 10\beta_1 + \beta_3 \neq 0,$$

[single linear combination of parameters]

$$bwgth = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 npvis + \beta_4 age + u,$$

- Are the partial effect of *age*, *education* and *npvis* jointly irrelevant after controlling for the number of cigarettes smoked?

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$
$$vs$$
$$H_1 : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0,$$

[jointly statistical significance; Exclusion restrictions]

- Is there any variable in the equation relevant to explain the birth weight?

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$vs$$
$$H_1 : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0,$$

[Overall significance of the regression]

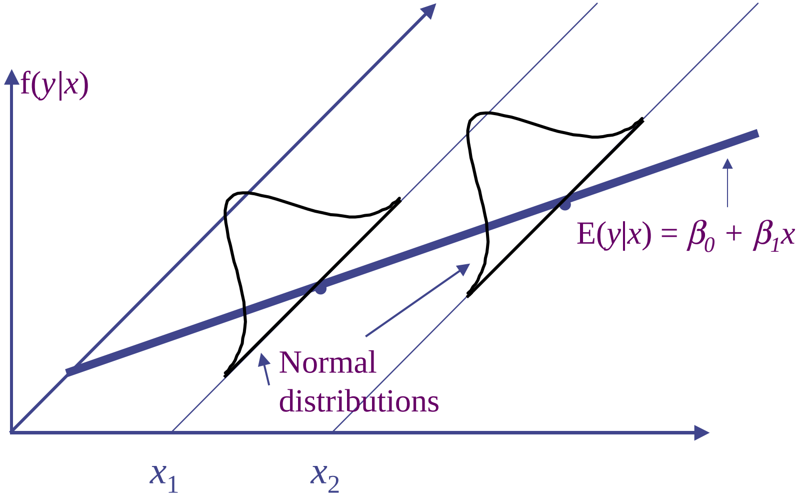
- So far, we know that given the Gauss-Markov assumptions, OLS is BLUE,
- In order to do classical hypothesis testing, we need to add another assumption (beyond the Gauss-Markov assumptions),
- Assume that u is independent of $x_1, x_2, \ldots, x_k$ and u is normally distributed with zero mean and variance σ^2: $u \sim N\left(0, \sigma^2\right)$.

- Under CLM, OLS is not only BLUE, but is the minimum variance unbiased estimator:
    - BLUE means that the OLS estimator is the most efficient among the class of linear unbiased estimators.
    - Under CLM the OLS estimator is the most efficient among all unbiased estimators.
- We can summarize the population assumptions of CLM as follows

$$y|x \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2).$$

- While for now we just assume normality, clear that sometimes not the case.
- Large samples will let us drop normality.

The homoskedastic normal distribution with a single explanatory variable



Under the CLM assumption, conditional on the sample values of the independent variables for $j = 0, ..., k$

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j)),$$

so that

$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1),$$

where $sd(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)}$.
$\hat{\beta}_j$ is distributed normally because it is a linear combination of independent errors that have the normal distribution.

Under the CLM assumptions

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t(n-k-1),$$

- Note this is a $t - student$ distribution because we estimate $sd(\hat{\beta}_j)$ by the standard error of $\hat{\beta}_j$, $se(\hat{\beta}_j)$,
- Note the degrees of freedom: $n-k-1$ (sample size-number of parameters of the model).
- In the simple regression model $k = 1$.

- Knowing the sampling distribution for the standardized estimator allows us to carry out hypothesis tests.
- Start with a null hypothesis $H_0 : \beta_j = b_j$, where b_j is a particular value.
- For example, $H_0 : \beta_j = 0$. If do not reject null, then x_j has no effect on the conditional mean of y, controlling for other x's.

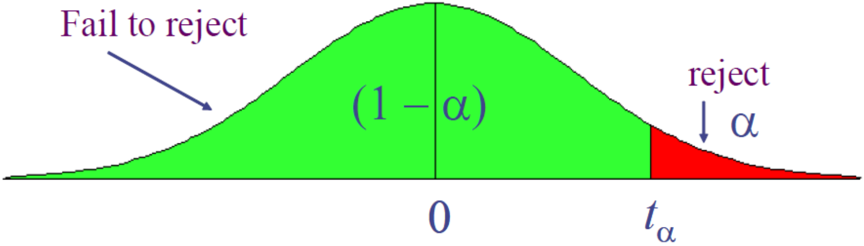
- To perform our test we first need to form the statistic : $t_j = \frac{\hat{\beta}_j - b_j}{se(\hat{\beta}_j)}$.
- Besides our null, H_0, we need an alternative hypothesis, H_1, and a significance level α.

Alternatives:

- $H_1 : \beta_j > b_j$ and $H_1 : \beta_j < b_j$ are one-sided.
- $H_1 : \beta_j \neq b_j$ is a two-sided alternative.

$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$

- $H_0 : \beta_j = b_j$ *vs* $H_1 : \beta_j > b_j$.
  Critical Value: t_α is defined as the constant that satisfies $\mathcal{P}(t_j > t_\alpha) = \alpha$, where t_j has the $t(n-k-1)$ distribution. Equivalently $\mathcal{P}(t_j < t_\alpha) = 1 - \alpha$.
  Rejection rule: Reject H_0 if the value of the t-statistic $> t_\alpha$.



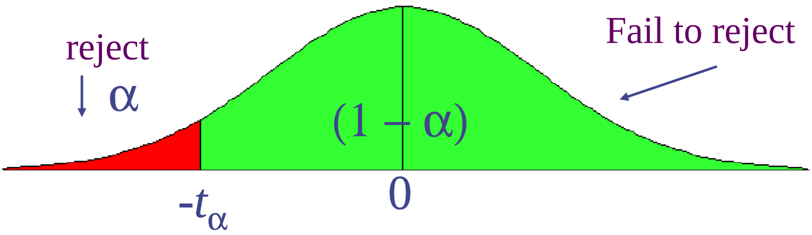
- **Example:** Consider the following regression where the standard errors are in brackets:

$$\begin{aligned}\widehat{\log(wages)} &= \underset{(0.104)}{0.284} + \underset{(0.007)}{0.092}educ + \underset{(0.0017)}{0.0041}exper + \underset{(0.003)}{0.022}tenure, \\ n &= 526,\ R^2 = 0.316\end{aligned}$$

Test whether, after controlling for education and tenure, higher work experience leads to higher hourly wages. Use the 5% and the 1% significance levels.

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

- $H_0 : \beta_j = b_j \; vs \; H_1 : \beta_j < b_j$.
  Critical Value: $-t_\alpha$ that is the constant that satisfies $\mathcal{P}(t_j < -t_\alpha) = \alpha$. where t_j has the $t(n-k-1)$ distribution.
  Equivalently $\mathcal{P}(t_j > -t_\alpha) = 1 - \alpha$.
  Rejection rule: Reject H_0 if the value of the t-statistic $< -t_\alpha$.



Example: Student performance and school size

- Consider the following regression

$$\begin{aligned}\widehat{math10} &= \underset{(6.113)}{2.274} + \underset{(0.0001)}{0.00046} totcomp + \underset{(0.04)}{0.048} staff - \underset{(0.00022)}{0.0002}\, enroll, \\ n &= 408, R^2 = 0.0541\end{aligned}$$

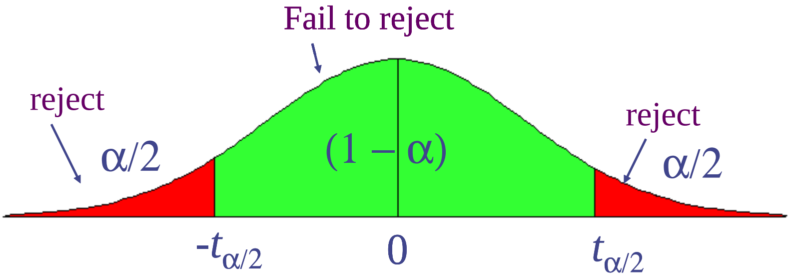
where

*math*10	-percentage of students passing math test
totcomp	-average annual teacher compensation
staff	-staff per one thousand students
enroll	-School enrollment=school size

Test whether smaller school size leads to better student performance at 5% level and 10% level.

$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$

- $H_0 : \beta_j = b_j$ *vs* $H_1 : \beta_j \neq b_j$.
  Critical Value: $t_{\alpha/2}$ is defined as the constant that satisfies $\mathcal{P}(t_j > t_{\alpha/2}) = \alpha/2$, where t_j has the $t(n-k-1)$ distribution.
  Rejection rule: Reject H_0 if the *absolute value* of the t-statistic $> t_{\alpha/2}$.



Example: Campus crime and enrollment
An interesting hypothesis is whether crime increases by one percent if enrollment is increased by one percent

$$\begin{aligned} \widehat{\log(crime)} &= \underset{(1.03)}{-6.63} + \underset{(0.11)}{1.27} \log(enroll), \\ n &= 97, R^2 = 0.0541 \end{aligned}$$

The estimate 1.27 is different from one but is this difference statistically significant? (use the 5% significance level)?

Remarks on $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$

- The quantity $t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ is called the t-ratio.
- If we reject the null, we typically say "x_j is statistically significant at the α level".
- If we fail to reject the null, we typically say "x_j is statistically insignificant at the α level".
- If asked to test whether a regressor is statistical significant, the alternative is assumed to be two-sided.

Example: Consider the following regression where the standard errors are in brackets:

$$\widehat{\log(wages)} = \underset{(0.104)}{0.284} + \underset{(0.007)}{0.092}educ + \underset{(0.0017)}{0.0041}exper + \underset{(0.003)}{0.022}tenure,$$
$$n = 526,\ R^2 = 0.316$$

Test whether, after controlling for experience and tenure, education is statistically significant at 5% and the 1% significance levels.

- The smallest significance level at which the null hypothesis is still rejected, is called the *p-value* of the hypothesis test
- A small p-value is evidence against the null hypothesis because one would reject the null hypothesis even at small significance levels
- A large p-value is evidence in favor of the null hypothesis

- Let t_j^{act} be the actual value of the t-statistic in the sample.
- If the alternative hypothesis is $H_1 : \beta_j > b_j$,

$$p - value = \mathcal{P}\left(t_j > t_j^{act}\right).$$

- If the alternative hypothesis is $H_1 : \beta_j < b_j$,

$$p - value = \mathcal{P}\left(t_j < t_j^{act}\right).$$

- If the alternative hypothesis is $H_1 : \beta_j \neq b_j$

$$p - value = \mathcal{P}\left(\left|t_j\right| > \left|t_j^{act}\right|\right).$$

- *Rejection rule*: If $p - value < \alpha$, we reject the null hypothesis.

Example: We are studying the returns to education at junior colleges and four year colleges (universities) and we have the model

$$\log(wages) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u,$$

where:

- jc =number of years attending a two year college
- $univ$ = number of years at a four year college
- $exper$ = months in workforce
- Data set taken from Kane and Rouse, 1995, “Labor Market Returns to Two- and Four-Year College", American Economic Review 85, 600-614. Sample size $n = 6,763$.

Running a regression of $\log\left(wages\right)$ on jc, $univ$ and $exper$ we obtain:

	Estimate	Std. Err.	t-Ratio	p-Value
Intercept	1.47233	0.02106	69.911	0
exper	0.00494	0.00016	30.901	0
jc	0.0667	0.00683	9.765	0
univ	0.07688	0.00231	33.28	0

$$n = 6763, R^2 = 0.2224.$$

This is the typical output of a software in a regression model. The p-value computed in this table is for the hypothesis $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$.

- Another way to use classical statistical testing is to construct a confidence interval using the same critical value as was used for a two-sided test.
- Using

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t(n-k-1),$$

we have

$$\mathcal{P}(\hat{\beta}_j - t_{\alpha/2} se(\hat{\beta}_j) < \beta < \hat{\beta}_j + t_{\alpha/2} se(\hat{\beta}_j)) = 1 - \alpha,$$

where $t_{\alpha/2}$ the constant that satisfies $\mathcal{P}(t_j < -t_{\alpha/2}) = \alpha/2$, where t_j is a random variable with distribution $t(n-k-1)$.
Equivalently $\mathcal{P}(t_j > t_{\alpha/2}) = \alpha/2$.

- Hence a $100(1-\alpha)\%$ confidence interval is defined as

$$(\hat{\beta}_j - t_{\alpha/2} se(\hat{\beta}_j), \hat{\beta}_j + t_{\alpha/2} se(\hat{\beta}_j)),$$

- In repeated samples, the interval that is constructed in the above way will cover the population regression coefficient in $100(1-\alpha)\%$ of the cases. The interval that we compute with the actual sample is one of these intervals
- Relationship between confidence interval and hypotheses tests:

$$b_j \quad \notin \quad conf.\,interval \Rightarrow \text{reject } H_0 : \beta_j = b_j \text{ in favour of } H_1 : \beta_j \neq b_j \text{ at } 100\alpha\% \text{ level.}$$

Example: Running a regression of $\log(wages)$ on *jc*, *univ* and *exper* we obtain:

	Estimate	Std. Err.	t-Ratio	p-Value
Intercept	1.47233	0.02106	69.911	0
exper	0.00494	0.00016	30.901	0
jc	0.0667	0.00683	9.765	0
univ	0.07688	0.00231	33.28	0

$$n = 6763, R^2 = 0.2224.$$

- Construct a 90% confidence interval for the coefficient of the variable *exper*.
- Construct a 95% confidence interval for the coefficient of the variable *jc*.
- Construct a 99% confidence interval for the coefficient of the variable *univ*.

- Suppose instead of testing whether β_1 is equal to a constant, you want to test if it is equal to another parameter, that is $H_0 : \beta_1 = \beta_2$.
- Use same basic procedure for forming a t statistic

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$

Notice that the standard error of $\hat{\beta}_1 - \hat{\beta}_2$, $se(\hat{\beta}_1 - \hat{\beta}_2)$, is an estimator of the standard deviation of $\hat{\beta}_1 - \hat{\beta}_2$:

$$\sqrt{Var\,(\hat{\beta}_1 - \hat{\beta}_2)}$$

Since

$$Var(\hat{\beta}_1 - \hat{\beta}_2) = Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2),$$

an estimator for $\sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)}$ is given by

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{se(\hat{\beta}_1)^2 + se(\hat{\beta}_2)^2 - 2s_{12}},$$

where s_{12} is an estimate of $Cov(\hat{\beta}_1, \hat{\beta}_2)$.

In some cases you can always restate the problem to get the test you want.
Example: Consider the model on the returns to education at junior colleges and four year colleges

$$\log(wages) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u,$$

- We would like to test whether one year at a junior college is worth one year at a university, that is $H_0 : \beta_1 = \beta_2$.
- The alternative hypothesis is that a year at junior college is worth less than a year at a university. That is $H_1 : \beta_1 < \beta_2$.
- One can test H_0 by using the approach described before.
- However there is an easier way.

Define a new parameter $\theta = \beta_1 - \beta_2$. Hence the null hypothesis becomes

$$H_0 : \theta = 0$$

and the alternative hypothesis becomes:

$$H_1 : \theta < 0,$$

We can always write the model in terms of θ. Under H_0, the model is equivalent to

$$\log(wages) = \beta_0 + \theta jc + \beta_2 totcoll + \beta_3 exper + u,$$

where $totcoll = jc + univ$.
This model is linear in the parameters so one can use the usual tests on hypothesis for single parameters described before.

- Running the regression of $\log(wages)$ on *exper*, *jc* and *totcoll* we obtain:

	Estimate	Std.Err.	t-ratio
Intercept	1.47233	0.02106	69.911
exper	0.00494	0.00016	30.901
jc	−0.01018	0.00694	−1.467
totcoll	0.07688	0.00231	33.28

$$n = 6763,\ R^2 = 0.2224$$

Test $H_0 : \theta = 0$ vs $H_1 : \theta < 0$ (use the 5% significance level).

Example (cont):

- This is the same model as originally, but now you get a standard error for $\hat{\beta}_1 - \hat{\beta}_2$ directly from the basic regression
- Any linear combination of parameters could be tested in a similar manner
- Other examples of hypotheses about a single linear combination of parameters: $\beta_1 = 1 + \beta_2$; $\beta_1 = 5\beta_2$; $\beta_1 = -(1/2)\beta_2$; *etc.*

- Everything we've done so far has involved testing a single linear restriction, (e.g. $\beta_1 = 0$ or $\beta_1 = \beta_2$)
- However, we may want to jointly test multiple hypotheses about our parameters.
- A typical example is testing "exclusion restrictions" – we want to know if a group of parameters are all equal to zero.

- Now the null hypothesis might be something like $H_0 : \beta_1 = 0, ..., \beta_q = 0$ in the model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q + ... + \beta_k x_k + u.$$

  That is, we want to test whether the parameters of the first q regressors (x_1 to x_q) are equal to zero.
- The alternative is just H_1: At least one of the $\beta_j \neq 0, j = 1, ..., q$.
- Can't just check each t statistic separately, because we want to know if the q parameters are jointly significant at a given level – it is possible for none to be individually significant at that level.

- To do the test we need to estimate the "restricted model" without $x_1, \ldots, x_q$ included, as well as the "unrestricted model" with all x's included and compute

$$F = \frac{\left(SSR_r - SSR_{ur}\right)/q}{SSR_{ur}/\left(n-k-1\right)},$$

  where SSR_r is the sum of squared residuals of the restricted model and SSR_{ur} is the sum of squared residuals of the unrestricted model.

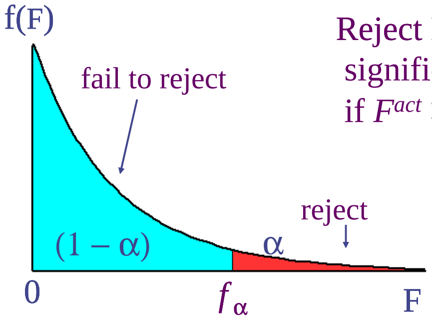
- Intuitively, we want to know if the change in SSR is big enough to warrant inclusion of $x_1, \ldots, x_q$.

- The F statistic is always positive, since the SSR from the restricted model can't be less than the SSR from the unrestricted.
- Essentially the F statistic is measuring the relative increase in SSR when moving from the unrestricted to restricted model.
- q =number of restrictions, or $df_r - df_{ur}$.
- $n - k - 1 = df_{ur}$.
- $n - k - 1 + q = df_r$.

- To decide if the increase in SSR when we move to a restricted model is "big enough" to reject the exclusions, we need to know about the sampling distribution of our F statistic.
- $F \sim F(q, n-k-1)$, where q is referred to as the numerator degrees of freedom and $n-k-1$ as the denominator degrees of freedom.
- Denote F^{act} the actual value of the statistic in a given sample.
- The critical value is denoted as f_α and corresponds to the constant that satisfies

$$\mathcal{P}(F > f_\alpha) = \alpha.$$

- *Rejection rule:* Reject H_0 if $F^{act} > f_\alpha$.



Reject H_0 at α significance level if $F^{act} > f_\alpha$

Example: Consider the following model that explains major league baseball players' salaries:

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

where

- *salary*= salary of major league baseball player
- *years* =Years in the league
- *gamesyr* =Average number of games per year
- *bavg* =Batting average
- *hrunsyr* =Home runs per year
- *rbisyr* =Runs batted in per year

We would like to test $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ vs $H_1 : H_0$ is not true.

- Estimating the unrestricted model we obtain

$$\begin{aligned}\widehat{\log(salary)} &= \underset{(0.29)}{11.19} + \underset{(0.0121)}{0.0689}years + \underset{(0.0026)}{0.0126}gamesyr \\ &\quad + \underset{(0.00110)}{0.00098}bavg + \underset{(0.0161)}{0.0144}hrunsyr + \underset{(0.0072)}{0.0108}rbisyr, \\ n &= 353, SSR = 183.186, R^2 = 0.6278\end{aligned}$$

- Estimating the restricted model we obtain

$$\begin{aligned}\widehat{\log(salary)} &= \underset{(0.11)}{11.22} + \underset{(0.0125)}{0.0713}years + \underset{(0.0013)}{0.0202}gamesyr, \\ n &= 353, SSR = 198.311, R^2 = 0.5971.\end{aligned}$$

- Test $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ vs $H_1 : H_0$ is not true at 5% level

- Because the SSR's may be large and unwieldy, an alternative form of the formula is useful.
- We use the fact that $SSR = SST(1 - R^2)$ for any regression, so can substitute in for SSR_r and SSR_{ur} :

$$F = \frac{\left(R_{ur}^2 - R_r^2\right)/q}{\left(1 - R_{ur}^2\right)/\left(n - k - 1\right)} \tag{1}$$

where R_r^2 is the R^2 of the restricted model and R_{ur}^2 is the R^2 of the unrestricted model.

Example: For the baseball salary example, use (1) to obtain the F statistic.

- A special case of exclusion restrictions is to test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0.$
- Since the R^2 from a model with only an intercept will be zero, the F statistic is simply

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

- **Example:** Consider the estimated model

$$\begin{aligned}\widehat{\log(salary)} &= \underset{(0.29)}{11.19} + \underset{(0.0121)}{0.0689}years + \underset{(0.0026)}{0.0126}gamesyr \\ &\quad + \underset{(0.00110)}{0.00098}bavg + \underset{(0.0161)}{0.0144}hrunsyr + \underset{(0.0072)}{0.0108}rbisyr, \\ n &= 353, SSR = 183.186, R^2 = 0.6278\end{aligned}$$

We would like to test

$$\begin{aligned} H_0 &: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0 \\ & \quad vs \\ H_1 &: H_0 \text{ not true} \end{aligned}$$

at 5% level.

- The basic form of the *F* statistic will work for any set of linear restrictions.
- First estimate the unrestricted model obtain SSR_{ur} and then estimate the restricted model and obtain SSR_r.
- The F statistic as the usual form

$$F = \frac{(SSR_r - SSR_{ur})\,/q}{SSR_{ur}/\,(n-k-1)} \sim F(q, n-k-1)$$

  where q is the number of restrictions being tested.
- Imposing the restrictions can be tricky – will likely have to redefine variables again.

Example: Test whether house price assessments are rational

$$\begin{aligned}\log(price) &= \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) \\ &\quad + \beta_3 \log(sqrft) + \beta_4 bdrms + u\end{aligned}$$

- *price* =Actual house price
- *assess* =The assessed housing value before the house was sold
- *lotsize* =Size of lot (in feet)
- *sqrft* =Square footage
- *bdrms* =number of bedrooms

- Now, suppose we would like to test whether the assessed housing price is a rational valuation. If this is the case, then a 1% change in assess should be associated with a 1% change in price; that is, $\beta_1 = 1$. In addition, *lotsize*, *sqrft*, and *bdrms* should not help to explain $\log(price)$, once the assessed value has been controlled for.
- Hence we want to test H_0: $\beta_1 = 1$, $\beta_2 = 0$, $\beta_3 = 0$, $\beta_4 = 0$ vs H_1: H_0 not true
- Sample size: 88.
- Running the regression of $\log(price)$ on $\log(assess)$, $\log(lotsize)$, $\log(sqrft)$ and *bdrms* we obtain $SSR_{ur} = 1.822$
- Imposing the restriction given by H_0 we have

$$\log(price) - \log(assess) = \beta_0 + u.$$

- Estimating the parameter of this model by OLS we obtain $SSR_r = 1.88$.
- Test H_0: $\beta_1 = 1$, $\beta_2 = 0$, $\beta_3 = 0$, $\beta_4 = 0$ vs H_1: H_0 not true at 5% level.

Prediction for the conditional mean of y

Suppose that we want an estimate of

$$E(y|x_1 = x_{1,0}, \ldots, x_k = x_{k,0}) = \beta_0 + \beta_1 x_{1,0} + \ldots + \beta_k x_{k,0}.$$

That is, we would like to estimate the the mean of y when the regressors are equal to known values $x_{1,0}, ..., x_{k,0}$.

- This is easy to obtain by substituting the x's in our estimated model with x_0's ,
$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \ldots + \hat{\beta}_k x_{k,0}.$$
- We would like to construct confidence intervals for $E(y|x_1 = x_{1,0}, \ldots, x_k = x_{k,0})$.
- But what about a standard error of $\hat{y}_0$, ?
- There is general formula for this standard error in the case $k > 1$, but it requires knowledge of matrix algebra. However there is a simple way to obtain this standard error.
- Let us change notation and define $\theta = E(y|x_1 = x_{1,0}, \ldots, x_k = x_{k,0})$.
- Thus now the objective becomes to construct a confidence interval for θ.
- θ is just a linear combination of the parameters.

Prediction for the conditional mean of y

- Can rewrite
$$\beta_0 + \beta_1 x_{1,0} + \ldots + \beta_k x_{k,0} = \theta$$
as
$$\beta_0 = \theta - \beta_1 x_{1,0} - \ldots - \beta_k x_{k,0}$$
- Substitute in
$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \; u \sim N(0, \sigma^2)$$
to obtain
$$y = \theta + \beta_1 (x_1 - x_{1,0}) + \ldots + \beta_k (x_k - x_{k,0}) + u$$
- So, if you regress y on $(x_j - x_{j,0}), j = 1, ..., k$, the intercept will give the predicted value and its standard error.
- Hence constructing a confidence interval for θ is similar to constructing a confidence interval for a parameter.
- $se(\hat{y}_0)$ is the standard error of the intercept in the regression of y on an intercept and $(x_j - x_{j,0}), j = 1, ..., k$.

Prediction for the conditional mean of y

Remark: In the simple regression model we have

$$y = \beta_0 + \beta_1 x + u,\ E(u|x) = 0,\ var(u|x) = \sigma^2$$

Suppose that we would like to predict the value of

$$E(y|x = x_0) = \beta_0 + \beta_1 x_0$$

In this case

$$se(\hat{y}_0)^2 = \hat{\sigma}^2 [\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}]$$

where $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{u}_i^2 / (n-2)$ (recall that $k = 1$ in the simple regression model).

Prediction for the conditional mean of y in the multiple regression model

Example: Consider the following equation:

$$y_i = \beta_1 + \beta_2 x_i + u_i, i = 1, ..., 60$$

The results from estimating this equation using 60 observations by Ordinary Least Squares were (standard errors in parentheses) are:

$$\hat{y} = \underset{(0.125)}{0.395} - \underset{(0.189)}{0.550}x,$$

$$SSR = 42.307, SSE = 6.1771,$$

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.34033$$

Given that $x_0 = 0.075$, the sample mean of x is 0.105 and that $u \sim N(0, \sigma^2)$, calculate the 95% confidence intervals for $E(y|x = x_0)$

Prediction for y

Suppose now that we would like to construct a confidence interval for y when when the regressors are equal to known values $x_{1,0}, ..., x_{k,0}$ and denote this value as y_0.

- How can we construct a confidence interval for y_0?
- Notice that

$$y_0 = \beta_0 + \beta_1 x_{1,0} + \ldots + \beta_k x_{k,0} + u_0$$

- Our best prediction for y_0 is the regression line

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \ldots + \hat{\beta}_k x_{k,0}$$

- The prediction error is given by

$$\begin{aligned} \hat{u}_0 &= y_0 - \hat{y}_0 \\ &= \beta_0 + \beta_1 x_{1,0} + \ldots + \beta_k x_{k,0} + u_0 - \hat{y}_0 \end{aligned}$$

- Therefore, as u_0 and $\hat{y}_0$ are independent (conditional on the regressors):

$$\begin{aligned} Var(\hat{u}_0) &= Var(u_0) + Var(\hat{y}_0) \\ &= \sigma^2 + Var(\hat{y}_0). \end{aligned}$$

Prediction for y

$$Var(\hat{u}_0) = \sigma^2 + Var(\hat{y}_0).$$

- Hence an estimator for $Var(\hat{u}_0)$ is given by

$$se_0^2 = \hat{\sigma}^2 + se(\hat{y}_0)^2,$$

  where $se(\hat{y}_0)$ is the standard error of the intercept in the regression of y on $(x_j - x_{j,0})$, $j = 1, ..., k$, and $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{u}_i^2 / (n - k - 1)$.
- It can be shown that if $u \sim N(0, \sigma^2)$,

$$\frac{y_0 - \hat{y}_0}{se_0} \sim t(n - k - 1)$$

- Hence the $(1 - \alpha)\%$ prediction interval for y_0 is given by

$$(\hat{y}_0 - t_{\alpha/2} se_0, \hat{y}_0 + t_{\alpha/2} se_0),$$

  where $t_{\alpha/2}$ is the percentile $(1 - \alpha/2)^{th}$ of the the t distribution with $n - k - 1$ df.

Prediction for y

Example: Suppose we have the following regression model

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &\quad + \beta_4 x_3^2 + u. \end{aligned}$$

We have a sample of 4,137 observations . The estimated model is

$$\begin{aligned} \hat{y} &= \underset{(0.075)}{1.493} + \underset{(0.00007)}{0.00149}\, x_1 - \underset{(0.00056)}{0.01386}\, x_2 - \underset{(0.01650)}{0.06088}\, x_3 \\ &\quad + \underset{(0.00227)}{0.00546}\, x_4, \\ \hat{\sigma} &= 0.56 \end{aligned}$$

Prediction for y

Objectives:

- Construct a 95% confidence interval for the mean of y when $x_1 = 1,200$, $x_2 = 30$ and $x_3 = 5$, $x_4 = 25$.
- Construct a 95% confidence interval for y when $x_1 = 1,200$, $x_2 = 30$, $x_3 = 5$, $x_4 = 25$.
- Define a new set of regressors:
  - $x_1^* = x_1 - 1,200$.
  - $x_2^* = x_2 - 30$.
  - $x_3^* = x_3 - 5$.
  - $x_4^* = x_4 - 25$.

Running the regression of y on these new regressors we obtain

$$\begin{aligned}\hat{y} &= \underset{(0.020)}{2.700} + \underset{(0.00007)}{0.00149}\, x_1^* - \underset{(0.00056)}{0.01386}\, x_2^* - \underset{(0.01650)}{0.06088}\, x_3^* \\ &\quad + \underset{(0.00227)}{0.00546}\, x_4^*. \\ \hat{\sigma} &= 0.56\end{aligned}$$

Predicting y in a log model

Suppose that we have the model

$$\log(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u,$$

$E(u|x_1, ..., x_k) = 0$, $Var(u|x_1, ..., x_k) = \sigma^2$ and we would like to predict the mean of y for any value of the regressors: $E(y|x_1, ..., x_k)$.
What can we do?
Given the OLS estimators the predicted value for the mean of $\log(y)$ for any values of the regressors is

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

Our first guess would be to exponentiate $\widehat{\log(y)}$.
However, simple exponentiation of $\widehat{\log(y)}$ will underestimate the expected value of y as $\widehat{\log(y)}$ is and estimator of $E(\log(y)|x_1, ..., x_k)$ and it can be shown using an inequality known as *Jensen's inequality* that

$$\exp[E(\log(y)|x_1, ..., x_k)] \leq E(y|x_1, ..., x_k).$$

Predicting y in a log model

If $u \sim N(0, \sigma^2)$, in can be shown that

$$E\left(y|x_1, ..., x_k\right) = \exp(\frac{\sigma^2}{2})\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k).$$

Therefore, a simple way to predict y is

$$\hat{y} = \exp(\frac{\hat{\sigma}^2}{2})\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k).$$