Lecture 5: Hypothesis Tests and Goodness of Fit

Thomas Chadefaux

Contents

1 Hypothesis testing 2

1.1 The t-test 2

1.2 What happens to our p-values if ... 3

1.3 Confidence Intervals for b 4

1.4 Confidence Intervals for y^0 5

1.5 Prediction Intervals 5

1.6 F test 6

2 Goodness of fit 8

2.1 R^2 8

2.2 AIC, BIC 8

1 Hypothesis testing

1.1 The t-test

To test whether we should include a variable or not, we can test its statistical significance. In particular, to check whether variable x_j has a statistically significant effect on y, we test the null hypothesis $H_0 : \beta_j = 0$ against the alternative $H_1 : \beta_j \neq 0$.

We can use the fact that estimator b is normally distributed with mean β and covariance matrix $\sigma^2(X'X)^{-1}$. First, let us normalise b by subtracting its mean and dividing by its standard deviation::

$$z = \frac{b_k - \beta_k}{\sigma\sqrt{c_{kk}}},$$

where c_{kk} is the (k,k) element of $(X'X)^{-1}$ (i.e., c_{kk} is the variance of coefficient b_k). z has a standard normal distribution (i.e., a normal distribution with mean 0 and variance 1. BUT since we do not know σ, we have to replace it by its sample estimate, s. This, however, implies that z is no longer distributed standard normally. But remember that

$$s^2 = \frac{e'e}{N-K'},$$

i.e. s^2 is the sum of squared normals, which implies that it is Chi-squared distributed (see lecture 1). Remember also from lecture 1 that if $Z \sim N(0,1), U \sim \chi^2_n$, and U and Z are independently distributed, then the variable

$$t = \frac{Z}{\sqrt{U/k}}$$

has a t distribution with k degrees of freedom.

Now note that we can define $Z = b_k - \beta_k$, which is normally distributed, $U = e'ec_{kk}$, which is chi-squared distributed, and $k = N - K$ degrees of freedom, so that

$$t_k = \frac{b_k - \beta_k}{s\sqrt{c_{kk}}} \sim t(N-K)$$

Note that with enough degrees of freedom (i.e., if you have enough observations), the t distribution becomes indistinguishable from the normal distribution.

Now that we know the distribution of b_k, we can finally construct test statistics and confidence intervals. In particular, we can test the hypothesis that a given coefficient is 0.

Derivation of t-test: First note that b is normally distributed. Why? Because it is a linear function of ε. Why? Remember that

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(Xb+\varepsilon) \\ &= AXb + A\varepsilon, \end{aligned}$$

where $A = (X'X)^{-1}X'$. So b is a linear function of ε. Since ε is normally distributed (by assumption), we find that b is also normally distributed. It has mean β and variance $\sigma^2(X'X)^{-1}$, so that

$$b \sim \mathrm{N}(\beta, \sigma^2(X'X)^{-1}).$$

```
setwd('~/Documents/Academia/Teaching/TCD/2015-HT/PO7005_Quantitative_Methods_II/Lectures/lecture5/')
# Set up the data
n <- 100
x <- rnorm(n)
Y <- x + 2*rnorm(n)

# Ask R to estimate the model
lm1 <- lm(Y ~ x)
summary(lm1)

#------- Let's do it all manually --------#
const <- rep(1, length(x))
X <- cbind(const, x)
XprimeXinv <- solve(t(X)%*%X)
XprimeY <- t(X)%*%Y
b = XprimeXinv %*% XprimeY

# find residuals
e <- Y - X%*% b

# sum of squared residuals:
eprimee <- t(e) %*% e

# estimated variance of residuals:
s2 <- as.numeric(eprimee / (length(x) -2) )

# variance of b:
var.b <- s2 *diag(2) %*% XprimeXinv
se.b <- sqrt(var.b)

# calculate t stat:
t <- b / diag(se.b)

# is this significantly different from 0?
# to do this manually (and approximately: this is just for intuition), generate many values drawn from a t-distribution with n-k df
random.t <- sort(rt(10000, df = n-2))
pdf('Figs/t_test.pdf')
hist(random.t, breaks=100, xlim=c(-max(abs(t)),max(abs(t))))
#draw a line at the minimum significance level, 97.5
xline(random.t[9750], lwd=2)
xline(random.t[250], lwd=2)
xline(t, col=2, lty=2)
axis(side = 1, at = t, labels = c('t0','t1'), cex.axis=2)
dev.off()
```

Listing 1: tstat.R

To do that, we know that

$$t_k = \frac{b_k - 0}{s\sqrt{c_{kk}}} = \frac{b_k}{se(b_k)} \sim t(N-K).$$

We reject the null hypothesis if the probability of observing a value of $|t_k|$ or larger is smaller than a given significance level α (typically 5%).

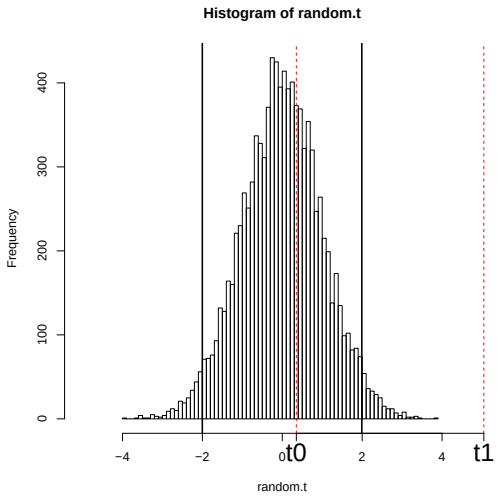


Figure 1: Figure: Plot of a t distribution with 2 degrees of freedom and t statistics of the regression coefficients (in red dotted lines). Source: tstat.R

If $p < 0.05$ (or whatever value you want), we say that b_k is significant at the .05 level, and typically give it a star in our regression output.

1.2 What happens to our p-values if ...

- Let's see what happens to t if we increase n: Then intuitively, our confidence in the estimate increases and hence our t value increases (and hence p-value decreases). More formally, $s^2 = \frac{e'e}{N-K}$ becomes smaller as N increases, and hence $t_k = \frac{b_k - \beta_k}{s\sqrt{c_{kk}}}$ is divided by an increasingly small number, which makes it become larger. A larger $|t|$ means a smaller p value.
- What happens to t if we increase the number of parameters to be estimated? K becomes larger, hence $N-K$ becomes smaller, hence s^2 becomes larger, hence t_k becomes smaller and our p value increases. In short, more parameters lead to higher standard errors and hence less significant coefficients.

A word of caution, however. Getting a low p-value (equivalently, a low s.e(b) or a high t) is nice, but make sure it is

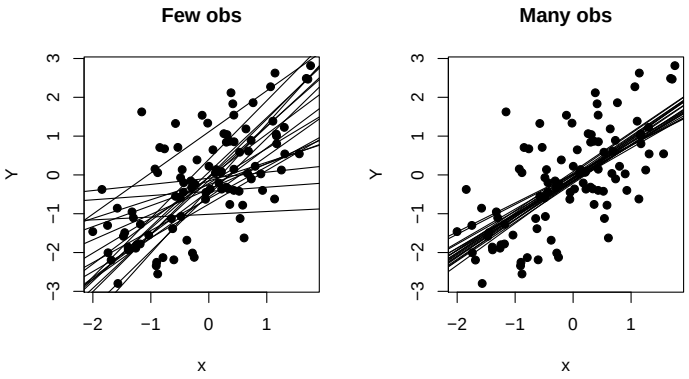


Figure 2: Changing the number of observations. Source: tstat.R

meaningful. In fact, given n sufficiently large, almost any coefficient will be significant. Here is an example:

Therefore you should ALWAYS look at the substantive interpretation of the coefficient. Is it different from 0 in a *meaningful* way?

1.3 Confidence Intervals for b

Now we'd like to use this information to make a claim about the probability that the interval $b_k \pm v$ contains β_k. First, remember from Stats 1 that

$$Prob[-1.96 \leq z_k \leq 1.96] = 0.95.$$

So by simple algebra, we get:

$$\begin{aligned} Prob[-1.96 \leq & z_k \leq 1.96 = 0.95 \\ Prob[-1.96 \leq & \frac{b_k - \beta_k}{\sigma\sqrt{c_{kk}}} \leq 1.96 = 0.95 \\ Prob[b_k - 1.96\sigma\sqrt{c_{kk}} \leq & \beta_k \leq b_k + 1.96\sigma\sqrt{c_{kk}}] = 0.95 \end{aligned} \quad (1)$$

IMPORTANT: 1.96 is an approximation based on the normal distribution. If you have very few observations, then it is important that you use the t distribution To see this, let us try an example. Consider the data $y = \begin{pmatrix} 10 \\ 12 \\ 3 \end{pmatrix}$ and $X = \begin{pmatrix} 1 & 4 \\ 1 & 9 \\ 1 & 1 \end{pmatrix}$. You want to estimate the following model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

You calculate $b = (X'X)^{-1}X'y = \begin{pmatrix} 3.43 \\ 1.05 \end{pmatrix}$ and $\hat{y} = Xb = \begin{pmatrix} 7.63 \\ 12.88 \\ 4.48 \end{pmatrix}$ and hence $e = \begin{pmatrix} -2.37 \\ 0.88 \\ 1.48 \end{pmatrix}$. The standard error of b, then, is:

$$\begin{aligned} var(b) = s^2(X'X)^{-1} &= \frac{e'e}{n-k}(X'X)^{-1} \\ &= 8.582 \times \begin{pmatrix} 1 & -0.1428 \\ -0.1428 & 0.03 \end{pmatrix} \end{aligned}$$

We are only interested in the square root of the diagonal, which gives us $se(b_0) = 2.93$ and $se(b_1) = 0.51$.

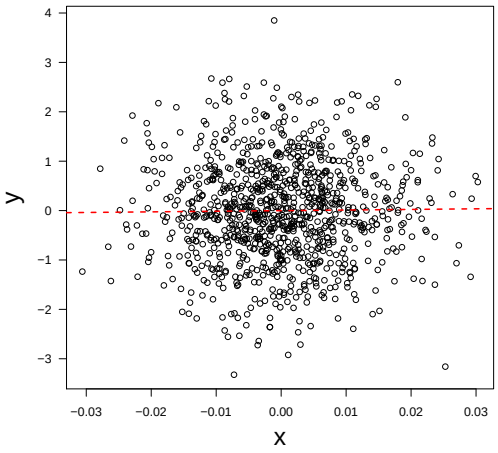


Figure 3: Figure: Why p-values are not everything: example of a regression line with $b_1 \approx 0$, but $p < 0.05$. See significance.R for source code.

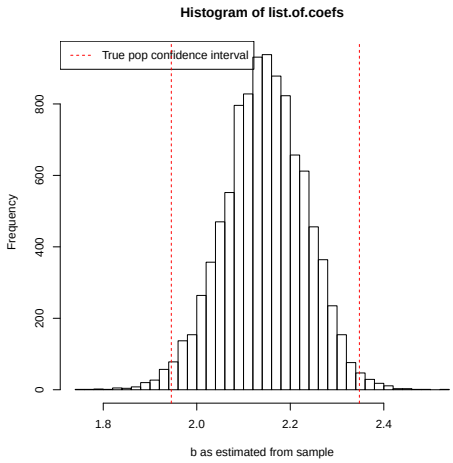


Figure: Suppose that we regressed y on x and calculated the confidence interval of b. We get, say, $b \in [1.88, 2.22]$. What does it mean? It means that if we drew a random sample from our population over and over, regressed y on x and collect b for each, 95% of the time we'd get $b \in [1.88, 2.22]$. This histogram shows that this is the case by plotting the distribution of the resulting bs. Source: ci_b.R

Finally we can get the confidence interval of b. Suppose we used 1.96 as our critical value for a 95% confidence interval. Then we'd obtain the following confidence interval for b_1:

$$1.05 \pm 1.96 \times 0.51 = [0.054, 2.065]$$

But actually we should use $t_{.0975,1} = 12.7$, giving us $1.05 \pm 12.7 \times 0.51 = [-5.42, 7.52]$. A huge difference!

1.4 Confidence Intervals for y^0

Now we want to calculate confidence bands around our predicted values. There are two types of uncertainty about our predicted values: uncertainty about the value of the coefficient, and uncertainty given the stochastic nature of the data (i.e., a different sample would yield different results). The *confidence interval* is a band around the fitted line, which tells us how confident we should be that the true coefficient is within these bands, given past values of the data (the ones we have at hand). However, it does not take into account the uncertainty associated with the data, . So if we want an confidence interval around *future* values, we need to take into account not only $Var[b]$ but also σ^2.

Let us start with the confidence interval around the value $\hat{y}^0$ associated with a regressor vector $\mathbf{x}^0$. To calculate it, we need to calculate $var[\hat{y}^0]$:

$$\begin{aligned} var[\hat{y}^0] &= Var[x^{0\prime} b] \\ &= x^{0\prime} Var[b] x^0 \\ &= x^{0\prime} \sigma^2 (X'X)^{-1} x^0 \end{aligned} \tag{2}$$

So we can write confidence interval is:

$$\text{confidence interval} = \hat{y}^0 \pm t_{0.975,n-K} se(\hat{y}^0)$$

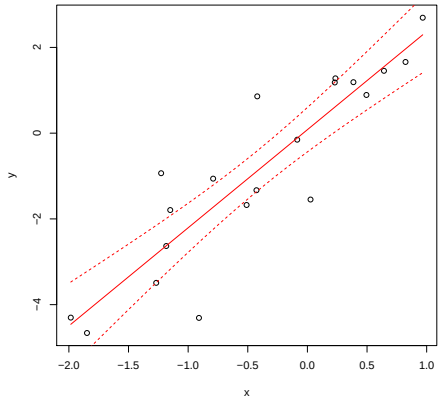
Intuitively, this is just the standard error of our coefficients multiplied by the vector of regressors $x^0 = \{1, x_1^0, x_2^0, \ldots, x_k^0\}$

1.5 Prediction Intervals

For a given x, say x^0, we want 1. to calculate $\hat{y}^0 = E[y|x^0]$, but also the confidence interval for y^0. We know that our error term for this value of x, e^0, is on average 0. But we need its variance in order to calculate a confidence interval.

How did we do step (2)?

$$\begin{aligned} Var(AX) &= E\left[(AX - E(AX))(AX - E(AX))'\right] \\ &= E[A(X - E(X))(X' - E(X)')A'] \\ &= E[A(X - E(X))(X - E(X))'A'] \\ &= AVar[X]A' \end{aligned}$$



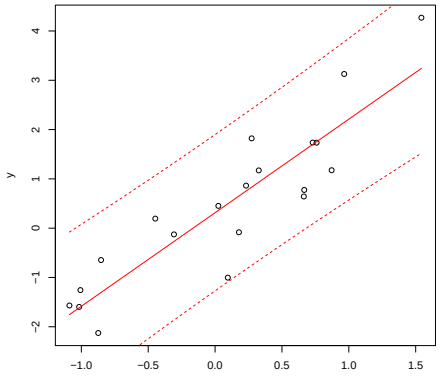


Figure 4: Left: Confidence Interval. Source: prediction_interval.R. Right: Prediction Interval. Notice how the bands are much wider than for the confidence interval. That is because the prediction interval also adds sampling uncertainty. Source: prediction_interval.R

Note first that

$$\begin{aligned} e^0 &= \hat{y}^0 - y^0 \\ &= (x^{0\prime} b) - (x^{0\prime}\beta + \varepsilon^0) \\ &= (b-\beta)' x^0 + \varepsilon^0 \end{aligned}$$

So

$$\begin{aligned} Var[e^0] &= Var[(b-\beta)' x^0 + \varepsilon^0] \\ &= Var[(b-\beta)' x^0] + \sigma^2 \end{aligned} \tag{3}$$

Now we note that $Var(AX) = AVar(X)A'$ if A is a constant matrix.

So (3) becomes:

$$Var[e^0] = \sigma^2 + x^{0\prime}[\sigma^2 (X'X)^{-1}] x^0$$

As usual, we need to replace σ^2 by s^2, but we now have the standard error of the prediction error, and hence can calculate a prediction interval as:

$$\text{prediction interval} = \hat{y}^0 \pm t_{0.975, n-K} se(e^0),$$

where $se(e^0) = \sqrt{Var[e^0]}$

1.6 F test

(from Verbeek:) A standard test that is often automatically supplied by a regression package as well is a test for the joint hypothesis that all coefficients, except the intercept β_1, are equal to zero. We shall discuss this procedure slightly more generally by testing the null that J of the K coefficients

That $Var[b-\beta] = Var[b]$ is left as an exercise.

are equal to zero ($J < K$). Without loss of generality, assume that these are the last J coefficients in the model,

$$H_0 : \beta_{K-J+1} = \beta_{K-J+2} = \beta_K = 0.$$

The alternative hypothesis in this case is that Ho is not true, i.e. that at least one of these J coefficients is not equal to zero. The easiest test procedure in this case is to compare the sum of squared residuals of the full model with the sum of squared residuals of the restricted model (which is the model with the last J regressors omitted). Denote the residual sum of squares of the full model by S1 and that of the restricted model by So. If the null hypothesis is correct one would expect that the sum of squares with the restriction imposed is only slightly larger than that in the unrestricted case

We want to compare the residual sum of squares of two models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (4)$$

$$y = \beta_0 + \beta_1 + \varepsilon \quad (5)$$

Let $RSS_{restricted}$ be the residual sum of squares of (5), and $RSS_{unrestricted}$ be the residual sum of squares (i.e., $\sum_i e_i$, aka $e'e$) of (4). Then the following statistic follows an F distribution with J and N-K degrees of freedom:

$$\frac{(RSS_{restricted} - RSS_{unrestricted})/J}{RSS_{unrestricted}/(N-K)}$$

Why do we care about F? Here are examples:

- First, suppose we are interested in who voted and who did not. We also have a series of dummy variables that measure whereas an individual identifies herself as Catholic, Protestant, Jewish, Muslim, or affiliated with another religion. We could run a regression with each dummy variable to see the rate at which each group votes. But the coefficients will always be in comparison to the omitted category, which may not be that useful. Most likely, we are interested in whether there is any difference between any of the groups. We can do that by testing the null hypothesis that all of the religion coefficients are equal to 0.
- We might also be interested in interaction terms (more on this in the next lecture). We can only rule them out if all of their constituent coefficients are equal to zero.

```
# Prediction Interval
n <- 20
x <- rnorm(n)
X <- cbind(rep(1,n), x)
#z <- rnorm(100)
y <- 2*x  + rnorm(n)

lm1 <- lm(y ~ 1 + x)

#--- canned solution
newx <- data.frame(x=seq(min(x), max(x), length.out=n))
pdf('Figs/pi.pdf')
plot(x, y)
ci1 <- predict(lm1, interval="predict", newdata = newx)
lines(newx$x, ci1[,1], lty=1, col=2)
lines(newx$x, ci1[,2], lty=2, col=2)
lines(newx$x, ci1[,3], lty=2, col=2)
dev.off()
# Calculate the ci for a specific value of x:
ci.canned <- predict(lm1, interval="predict", newdata = data.frame(x=x
      [1]), se.fit=T)

#--- let's get the same thing manually:
xprimex <- t(X)%*%X
b <- solve(xprimex) %*% t(X)%*%y
yhat <- X%*%b
e <- yhat - y
s2 <- as.numeric(t(e)%*%e) / (n-2)
var.b <- s2 * solve(xprimex)
se.e0 <- sqrt(s2 + t(X[1,])%*%var.b%*%X[1,])
t <- qt(0.975, n-2)
ci.home.made <- c(yhat[1] -t*se.e0, yhat[1] +t*se.e0)
#same result as in ci.canned


# Note: how about the CONFIDENCE interval?
se <- sqrt(t(X[1,])%*%var.b%*%X[1,])
t <- qt(0.975, n-2)
ci.home.made <- c(yhat[1] -t*se, yhat[1] +t*se)
ci.canned <- predict(lm1, interval="confidence", newdata = data.frame(x=x
      [1]), se.fit=T)
pdf('Figs/ci.pdf')
plot(x, y)
ci1 <- predict(lm1, interval="confidence", newdata = newx)
lines(newx$x, ci1[,1], lty=1, col=2)
lines(newx$x, ci1[,2], lty=2, col=2)
lines(newx$x, ci1[,3], lty=2, col=2)
dev.off()
```

Listing 2: prediction_interval.R

An example: Suppose $y = \begin{pmatrix} 5 \\ 7 \\ 12 \\ 2 \end{pmatrix}$ and $X = \begin{pmatrix} 1 & 3 & 5 \\ 1 & 2 & 9 \\ 1 & 5 & 1 \\ 1 & 3 & 2 \end{pmatrix}$ The sum of squared residuals for the model $y = \beta_0 + \beta_1 x + \varepsilon$ is 1.917391 (in R: sum(resid(lm1)^2)) The sum of squared residuals for the model $y = \beta_0 + \varepsilon$ is 53. The question then is how bad is 53 compared to 1.91? Let us calculate the F statistic:

$$\frac{(53 - 1.91)/2}{1.91/(4-3)} = 13.32$$

Now we look up 2.008 in a table of an F distribution with 2 and 1 degrees of freedom and get a p-value of 0.1902. I.e., we cannot be confident that at least one coefficient is not equal to 0.

- Policies might be significant, but only together.
- Your F statistic might be significant, yet none of your coefficient is statistically different from 0. This could be because of multicollinearity (more on this later).

2 Goodness of fit

Now that we have estimated the coefficient parameters, we want to ask how well the estimated regression line fits the observations.

2.1 R^2

One way to do this is to calculate the sum of squared residuals. I.e., for each observation, we calculate $\hat{y} - y$, square it, and sum these up. But we would like a more meaningful estimate of how good our line is fitting the points. In particular, we want to know what % of the variation in y is explained by the line.

First let's think about the total variation in y. That's the sum of squared deviations from the mean, i.e., $\sum_i (y_i - \bar{y})^2$. Now, we know that some of the variation is NOT explained by our line. What portion? Our error term. That is, $\sum_i (\hat{y}_i - y)^2 = e_i^2$. So the portion of the variation in y that is not explained by our regression line is:

$$\frac{e'e}{\sum_i (y_i - \bar{y})^2}$$

And since we want to know how much IS explained by our regression line, we subtract this from 1. I.e.,

$$R^2 = 1 - \frac{e'e}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{Var[e]}{Var[y]} = \frac{\text{Regression variation}}{\text{Total Variation}}$$

R^2 should almost never be your guide to select your regression. First, R^2 never decreases as you add variables, so it would be tempting to keep adding them. Remember, however, that this is at the cost of lower t values and hence less significant coefficients.

2.2 AIC, BIC

There are other measures of fit that penalises a large number of parameters more heavily. In particular, AIC and BIC

```r
# R2
x <- rnorm(100)
y <- x + rnorm(100)
lm1 <- lm(y ~ x)

total.var.in.y <- sum((y - mean(y))^2)
variation.not.explained.by.x <- sum(residuals(lm1)^2)
R2 <- 1-( variation.not.explained.by.x / total.var.in.y)
R2

#let's check:
summary(lm1)$r.squared
```

Listing 3: R2.R

are important:

$$AIC = log\frac{1}{N}e'e + \frac{2K}{N}$$
$$BIC = log\frac{1}{N}e'e + \frac{K}{N}log(N)$$

Models with a *lower* AIC or BIC are preferred.