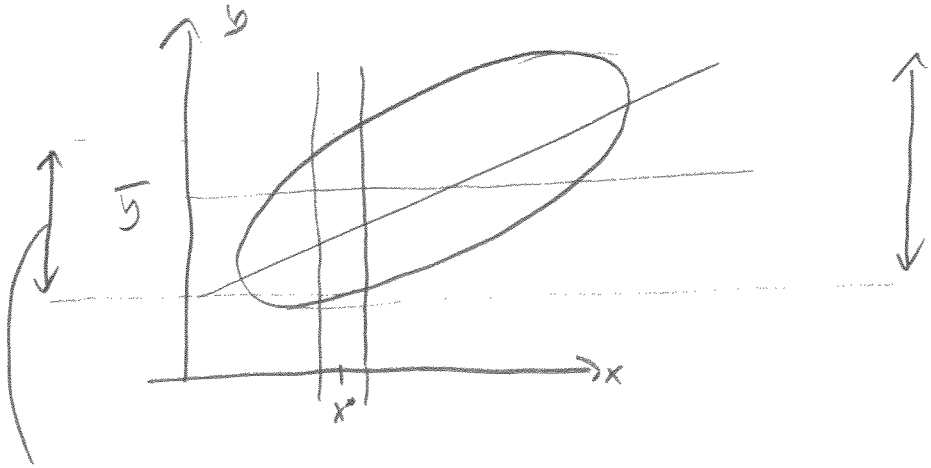
HANDOUT 4

Variance Decomposition



Spread in y ~ V(y)
marginal variance

Spread in y when x in neighborhood of x^*
Spread around regression line

$V(y|x=x^*)$

(conditional variance

(We assume $V(y|x) = \sigma^2$ constant, i.e. spread around the line does not depend on x value)

If $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$

$V(y)$ = Spread around $\bar{y}$ >> σ^2 if β_1 large

$V(y|x) = V(\varepsilon) = \sigma^2$

Note, if y and x are unrelated ($\beta_1 = 0$) $\Rightarrow V(y) = V(y|x)$
$\Rightarrow$ regression meaningless

If $y = \beta_0 + \beta_1 x$ exactly, $V(y|x) = 0$

$\Rightarrow$ all variability in y explained by x

Comparing marginal & conditional variability — are y and x related?

(1) RSS = Residual sum of squares = SS_E = error sum of squares

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

Spread around the regression line

$\sim V(y|x)$

(2) SS_T - Total sum of squares = $\sum_{i=1}^{n} (y_i - \bar{y})^2$

Spread around mean of y.

$\sim V(y)$

Now, since we fit the line to the data to minimize $Q \Rightarrow$ RSS smallest possible Q-value

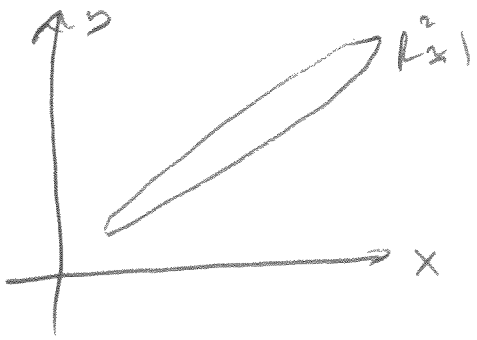
whereas $SS_T = Q(\beta_1 = 0)$

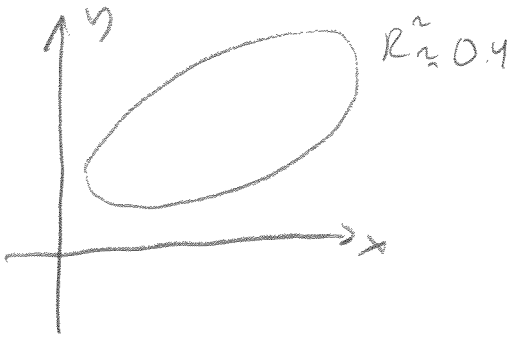
$\Rightarrow$ RSS always smaller than SS_T

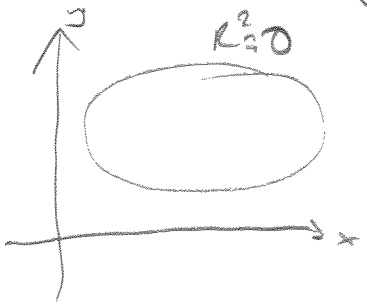
How much smaller should it be for regression to be meaningful?

$$R^2 = \frac{SS_T - RSS}{SS_T}$$ Compares $\frac{\text{The reduction in spread}}{\text{total spread}}$

= % of variability in y explained by regression on x.







Now, the reduction $SS_T - RSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

can be re-expressed as

$$\Sigma(y_i - \bar{y})^2 - \Sigma(y_i - \hat{y}_i)^2$$

($\hat{y}_i + e_i$)

$$= \cancel{\Sigma y_i^2} + \Sigma \bar{y}^2 - 2\bar{y}\Sigma y_i - \cancel{\Sigma y_i^2} - \Sigma \hat{y}_i^2 + 2\Sigma y_i \hat{y}_i$$

$\Sigma \bar{y}^2 = -n\bar{y}^2$

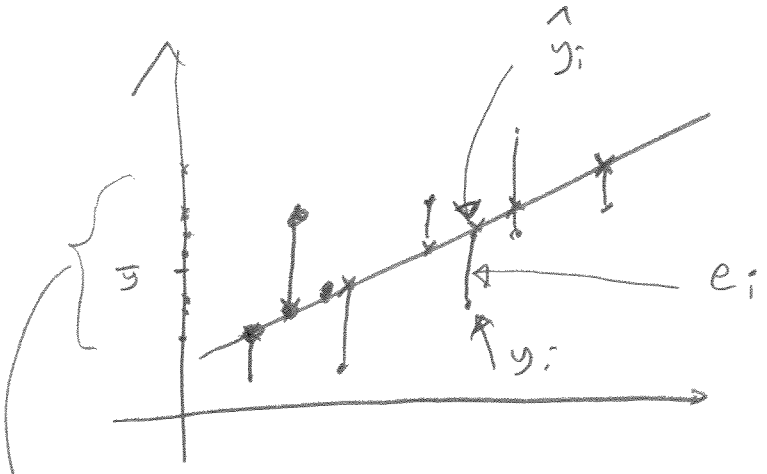
$$= -n\bar{y}^2 - \Sigma \hat{y}_i^2 + 2\Sigma \hat{y}_i^2 + 2\underbrace{\Sigma e_i \hat{y}_i}_{0}$$

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

= spread of fitted values around mean of y

= regression sum of squares

$= SS_{reg}$



spread of $\hat{y}$'s around $\bar{y}$

So, $$SS_T = RSS + SS_{reg}$$ Variance decomposition

SS_T: Total sum of squares $\sim V(y)$

RSS: residual sum of squares $\sim V(y|x)$

SS_{reg}: regression sum of squares — spread induced in y-values through the model only — no noise.

Note, $R^2 = \frac{SS_T - RSS}{SS_T} = \frac{SS_{reg}}{SS_T}$ = % of variability in y accounted for by the model.

Formally deciding if y and x are related

→ the F-test

(Goodness of fit or Lack-of-fit test)

We want to compare how much variability is explained by the model (SS_{reg}) & how much is left (RSS) — what level of % variability explained is enough to conclude y and x are related?

We will work out how much RSS and SS_{reg} can differ just by chance, i.e. when y and x are actually unrelated

Null hypothesis: y and x are unrelated, i.e. true $\beta_1 = 0$

① If $\beta_1 = 0 \Rightarrow y_i = \beta_0 + \varepsilon_i$, $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$

$$\Rightarrow SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

If null is true $\mathbb{E}_{\beta_1=0}(SS_T) = (n-1)\sigma^2$

That is, if null is true, we only need $\hat{\beta}_0 = \bar{y}$ in the model, so SS_T gives an estimate of the noise σ^2: $\hat{\sigma}^2 = \frac{SS_T}{n-1}$

(2) On the other hand, last lecture we showed that $\mathbb{E}(RSS) = (n-2)\sigma^2$

We can fit a line to the data even if true $\beta_1 = 0$ and the RSS also gives an estimate of σ^2

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

(3) Also, <u>if</u> the null is true, turns out there is a third estimate of σ^2.

$$\mathbb{E}_{\beta_1=0}(SSreg) = \mathbb{E}_{\beta_1=0}\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right)$$

$$= \mathbb{E}_{\beta_1=0}\left(\sum \hat{y}_i^2 + \bar{y}^2 - 2\bar{y}\hat{y}_i\right)$$

$$= \mathbb{E}_{\beta_1=0}\left(\sum \hat{y}_i^2\right) + n\,\mathbb{E}_{\beta_1=0}(\bar{y}^2) - 2\,\mathbb{E}_{\beta_1=0}\left(\bar{y}\sum \hat{y}_i\right)$$

$$= \sum_{i=1}^{n}\left[V(\hat{y}_i) + (E(\hat{y}_i))^2\right] + n\,V(\bar{y}) + n(E(\bar{y}))^2$$

$$- 2\sum_{i=1}^{n}\left[\mathrm{Cov}(\bar{y}, \hat{y}_i) + E(\bar{y})E(\hat{y}_i)\right]$$

$$= \sigma^2\sum_{i=1}^{n} h_{ii} + n\beta_0^2 + n\frac{\sigma^2}{n} + n\beta_0^2 - 2\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\mathrm{Cov}\left(\frac{y_j}{n}, \hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j\right)\right] - 2n\beta_0^2$$

$$= \sigma^2\sum_{i=1}^{n} h_{ii} + \sigma^2 - 2\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{\sigma^2}{n}h_{ij}\right) = \sigma^2\sum_{i=1}^{n} h_{ii} + \sigma^2 - \frac{2\sigma^2}{n}\sum_{i=1}^{n}\underbrace{\sum_{j=1}^{n} h_{ij}}_{=1}$$

$$= \sigma^2\sum_{i=1}^{n} h_{ii} + \sigma^2 - 2\sigma^2 = \sigma^2\left(\sum_{i=1}^{n} h_{ii} - 1\right)$$

Tricks

- $V(z) = E(z^2) - (E(z))^2$
  $\Rightarrow E(z^2) = V(z) + (E(z))^2$
- $\mathrm{Cov}(z, w) = E(zw) - E(z)E(w)$
  $\Rightarrow E(zw) = \mathrm{Cov}(z, w) + E(z)E(w)$
- If null is true, $\beta_1 = 0$
  $E(y_i) = \beta_0$
  $E(\hat{y}_i) = \beta_0$

So, under the null (if $\beta_1 = 0$)

$$\mathbb{E}_{\beta_1=0}(SS_{reg}) = \sigma^2 \left(\sum_{i=1}^{n} h_{ii} - 1 \right)$$

$$= \sigma^2 \qquad \left(\sum_{i=1}^{n} h_{ii} = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} \right) = 2 \right)$$

Note, if we do multivariate regression, i.e. we fit

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

model to the data, there are $\underline{\underline{p}}$ model parameters:

$\{\beta_0, \beta_1, \ldots, \beta_{p-1}\}$ ($\Rightarrow$ note $p = 2$ in simple regression.)

Then, $\mathbb{E}(RSS) = \sigma^2 (n-p)$ so $\hat{\sigma}^2 = \frac{RSS}{n-p}$

Under null $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ (all slopes are 0)

$$\mathbb{E}(SS_{reg}) = \sigma^2 (p-1)$$

<u>OK</u> So if null is true we have 3 estimate of σ^2 to compare

(1) $\frac{SST}{n-1}$ (only works if null is true)

(2) $\frac{RSS}{n-p}$ (works if null is true or not)

(3) $\frac{SS_{reg}}{p-1}$ (only works if null is true)

If null is not true ① and ③ will be inflated compared with ②.

We choose to compare ② and ③ (not ①).

Why? Because ② $= \frac{RSS}{n-p}$ = sum of squares of residuals / $n-p$

and ③ $= \frac{SSreg}{p-1}$ = spread among fitted values $\hat{y}$

and we know residuals e and fitted values $\hat{y}$ are uncorrelated $\Rightarrow$ easier to work out the distribution for ③/②.

Extra assumption needed $\varepsilon_i \sim$ normally distributed

$$\Downarrow$$

$$\frac{SSreg}{\sigma^2} \sim \chi^2_{p-1} \quad \text{Independent} \quad \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$$

and definition of F-distribution

$$\frac{\chi^2_{v}/v}{\chi^2_{n}/n} = F_{v,n}$$

Test statistic

$$F_{observed} = \frac{SSreg / p-1}{RSS / n-p}$$

$$= \frac{(SS_T - RSS) / ((n-1) - (n-p))}{RSS / n-p}$$

$$= \frac{\text{reduction in spread} / \text{difference in nbr of parameters}}{\text{Mean Squared Error}}$$

If null is true ($\beta_1 = 0$)

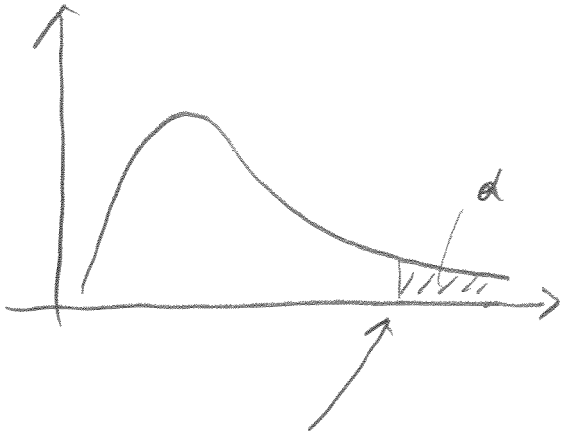
$F_{observed} \simeq 1$ since (2) and (3) both estimate σ^2

If null is not true ($\beta_1 \neq 0$)

$F_{observed} >> 1$.

How much bigger than 1 should F be for us to reject the null?

Compare $F_{observed}$ to the F-distribution $F_{p-1, n-p}$

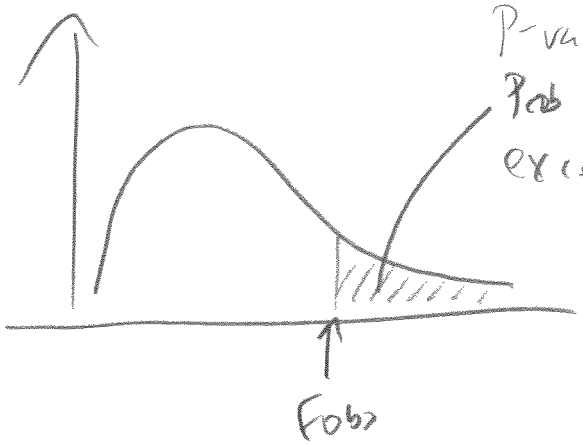


Histogram of the $F_{p-1, n-p}$ distribution

Look for the critical value F^* where according to $F_{p-1, n-p}$ only α % of F-values exceed this value.

Now, if $F_{obs} > F^*$ it is unlikely that the data comes from the null since only α % of the time would you be so 'unlucky'. $\Rightarrow$ Reject null at the level α.

Or, compute the p-value



P-value = Prob that values from $F_{p-1, n-p}$ exceed observed F_{obs}.

this large

If p-value is small, unlikely to get F_{obs} by chance alone $\rightarrow$ reject the null

Inference about β_1

Again, assume $\varepsilon_i \sim N(0, \sigma^2)$

$\Rightarrow \hat{\beta}_1 = \sum_{i=1}^{n} k_i \cdot y_i$, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

= linear combination of y-values, y ~ normally distributed

$\Rightarrow \hat{\beta}_1$ normally distributed

Already know $E(\hat{\beta}_1) = \beta_1$, $V(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$

$$\Rightarrow \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}\right)$$

or

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}}} \sim N(0,1)$$

but we don't know σ^2

We estimate σ^2 as $\hat{\sigma}^2 = \dfrac{RSS}{n-p}$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}} \neq N(0,1)$$

Instead $\hat{\beta}_1 \sim$ normally distributed

$\hat{\sigma}^2 \propto RSS \sim$ sum of errors squared

= sum of normally distributed variables squared

$\sim \chi^2$ distributed

Definition $\frac{N}{\chi^2} =$ t-distribution

So here,

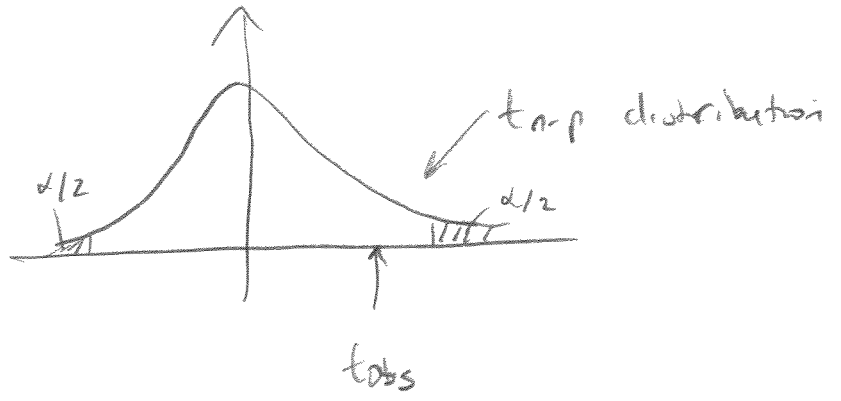
$$t \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}} \sim t_{n-p}$$

degrees of freedom

How to use?

① Testing Compute $t_{obs} = \dfrac{\hat{\beta}_1 - 0}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}}$ ← assumed null value

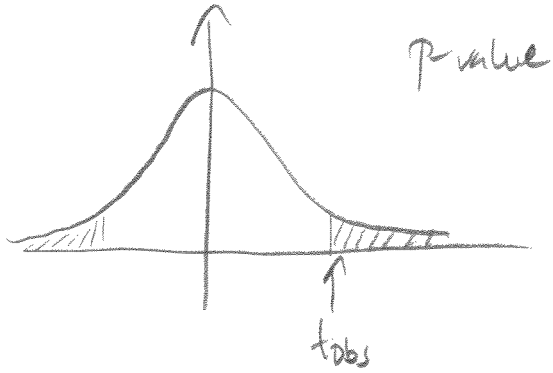
and compare with distribution t_{n-p}



If $|t_{obs}|$ exceeds critical value $t_{n-p}(1-\alpha/2)$

→ reject the null at level α

Alternatively, compute p-value



p-value = //// region

= amount of probability mass in t_{n-p} distribution at values that exceed t_{obs} (or are below $-t_{obs}$)

(2) Confidence Intervals

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}} \sim t_{n-p}$$

$1-\frac{\alpha}{2}$ percentile of t_{n-p} distribution

$$\Rightarrow P\left(\frac{|\hat{\beta}_1 - \beta_1|}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}} \leq t_{n-p}(1-\alpha/2) \right) = 1-\alpha$$

$$\Rightarrow P\left(-t_{n-p}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}} \leq \hat{\beta}_1 - \beta_1 \leq t_{n-p}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}} \right)$$

$$\Rightarrow P\left(\beta_1 - t_{n-p}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}} \leq \hat{\beta}_1 \leq \beta_1 + t_{n-p}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}} \right) = 1-\alpha$$

That is, $\hat{\beta}_1$ is highly likely to fall in the neighborhood of β_1, size of neighborhood $= \pm t_{n-p}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}$

Can rewrite as

$$P\left(\hat{\beta}_1 - t_{n-p}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}} \le \beta_1 \le \hat{\beta}_1 + t_{n-p}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}} \right) = 1$$

meaning the random interval

$$\left[\hat{\beta}_1 \pm t_{n-p}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}} \right] \quad (*)$$

is highly likely to cover the true value β_1.

Using interval for testing:

If $(*)$ does not cover 0, reject the null hypothesis.

Report writing

Structure

(1) Introduction

- State GOALS
- Maybe a preview of the most important result

(2) Results

- Summarize the analysis
- Use subsections e.g. . Data Transformations
  . Least Squares fit
  . Diagnostics
  . Outliers
  etc

(3) Conclusion

- Interpret results
- Most important discovery
- Problems or surprises?
- What is the next step? (what you would have done next)

(4) Appendix

To avoid putting too much emphasis on necessary but not so enlightening work - put this in appendix. In later labs, basic stuff like plotting data, [illegible] will go here.

General

- Use full sentences in your report.
- Don't mix fonts or fontsizes too much.
- Don't use cut-and-paste from software output → put this into tables instead
- Tables and Figures need to be numbered and captioned.
- Caption = description of table content, figure content + 1-7 sentences that summarizes the "message"
- Use subsections, subsubsections, paragraphs to avoid having too dense of a report
- Spell check