Stat 135 Dec 19

Today Finish Regression (Sec 14.2)

Thursday Bayesian Stats (Sec 8.6)

Monday Review with Adam 9-11 AM here

Tuesday no class. GSI extended OH.

Wednesday Final 9-12 here.

Simple Linear regression (sec 14.2)

Last time

95% Confidence interval for B_1 is

$$\hat{B}_1 \pm t_{n-2}(0.025)\, S_{\hat{B}_1}$$

where $S_{\hat{B}_1} = \frac{s^2}{n\,\mathrm{var}(x)} = \frac{RSS}{(n-2)\,n\,\mathrm{var}(x)}$.

SE of regression line

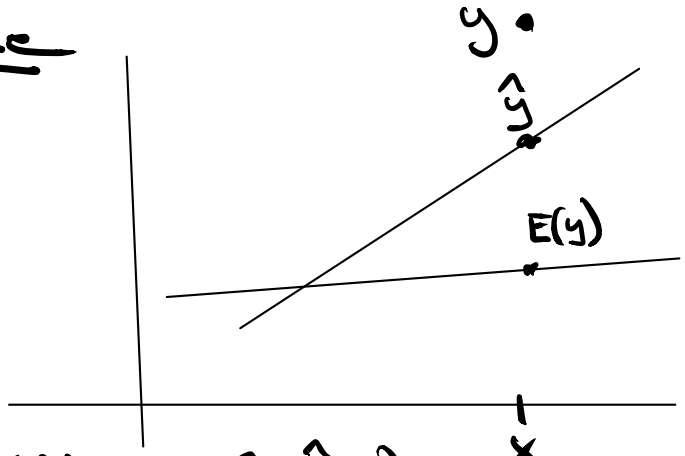
Let x be a pt on x axis.

If y is paired with x, $y = B_0 + B_1 x + e$

$\Rightarrow E(y) = B_0 + B_1 x$ so $E(y)$ lies on Tyche line.

Denote x on regression line by $\hat{y}$

Picture



regression line
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

true line
$E(y) = \beta_0 + \beta_1 x$

$E(\hat{y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x) = \beta_0 + \beta_1 x$ so $\hat{y}$ is an unbiased est of $E(y)$.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x$$

$$\boxed{\hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x})}$$

$$\text{var}(\hat{y}) = \text{var}\left(\bar{y} + \hat{\beta}_1 (x - \bar{x})\right)$$

$$= \text{var}(\bar{y}) + (x - \bar{x})^2 \text{var}(\hat{\beta}_1) + 2(x - \bar{x}) \text{cov}(\bar{y}, \hat{\beta}_1)$$

$\text{cov}(\bar{y}, \hat{\beta}_1) = 0$

Show $\text{cov}(\bar{y}, \hat{\beta}_1) = 0$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 n \bar{x}}{n^2 \text{var}(x)}$$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\text{cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) = \text{cov}(\bar{y}, \hat{\beta}_1) - \bar{x} \text{var}(\hat{\beta}_1)$$

$\text{var}(\hat{\beta}_1) = \frac{n\sigma^2}{n^2 \text{var}(x)}$

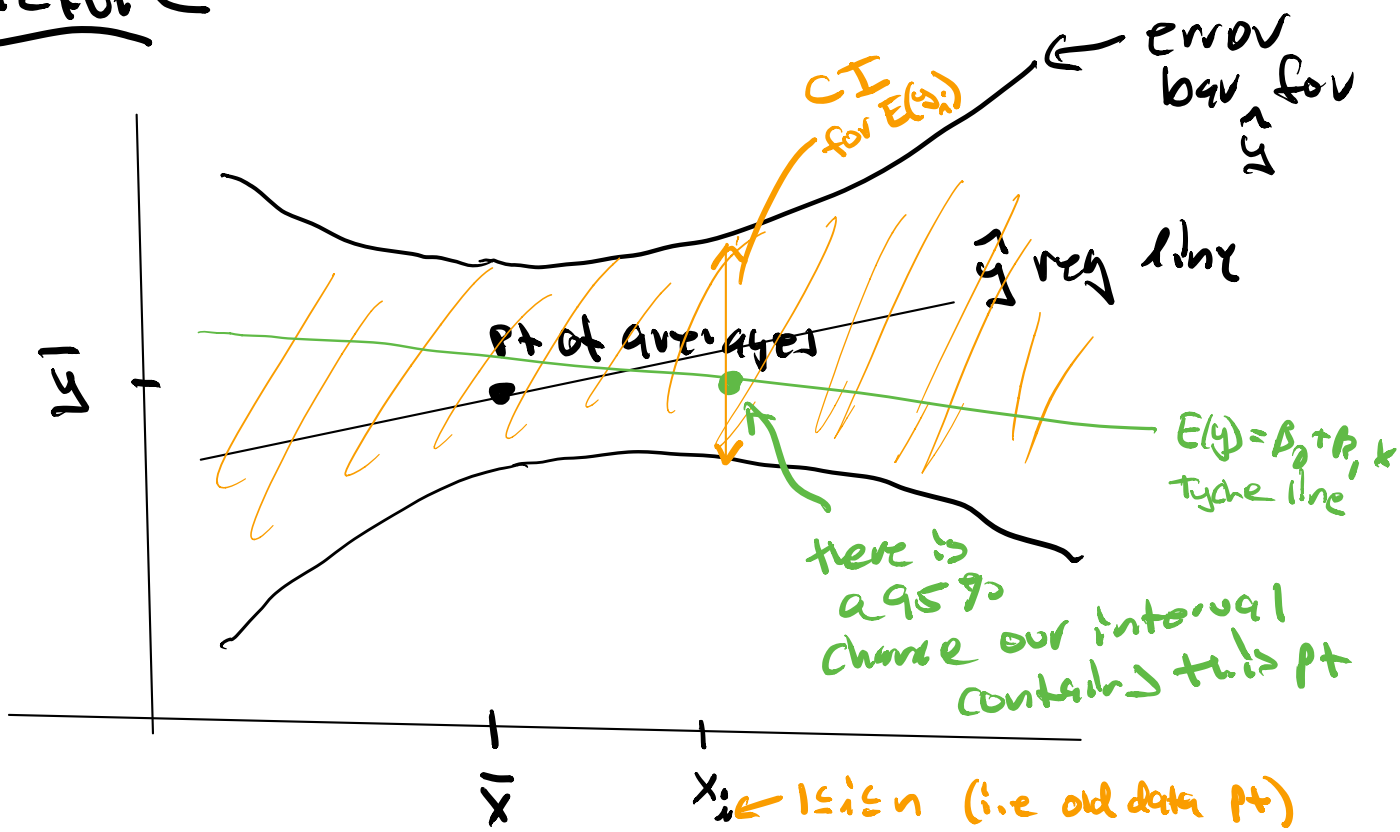
$\Rightarrow \text{cov}(\bar{y}, \hat{\beta}_1) = 0$ ✓

so $\text{var}(\hat{y}) = \frac{\sigma^2}{n} + (x-\bar{x})^2 \frac{\sigma^2}{n\,\text{var}(x)}$

so $\text{var}(\hat{y}) = \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{n\,\text{var}(x)} \right]$

$\sigma^2 \approx s^2$

Picture



Confidence interval for Tyche line $E(y)$

mean of y ← $\beta_0 + \beta_1 x$

A 95% CI for $E(y) = \beta_0 + \beta_1 x$

is $\hat{y} \pm t_{n-2}(0.025)\, s_{\hat{y}}$ where

$$s^2_{\hat{y}} = s^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{n\,\text{var}(x)} \right)$$

Prediction interval (PI) for a new observation y_{n+1}. — not in book.

A confidence interval for a RV is called a prediction interval.

A new observation y_{n+1} is a RV.

You give x_{n+1} and you get y_{n+1}.

Find a 95% PI for y_{n+1}.

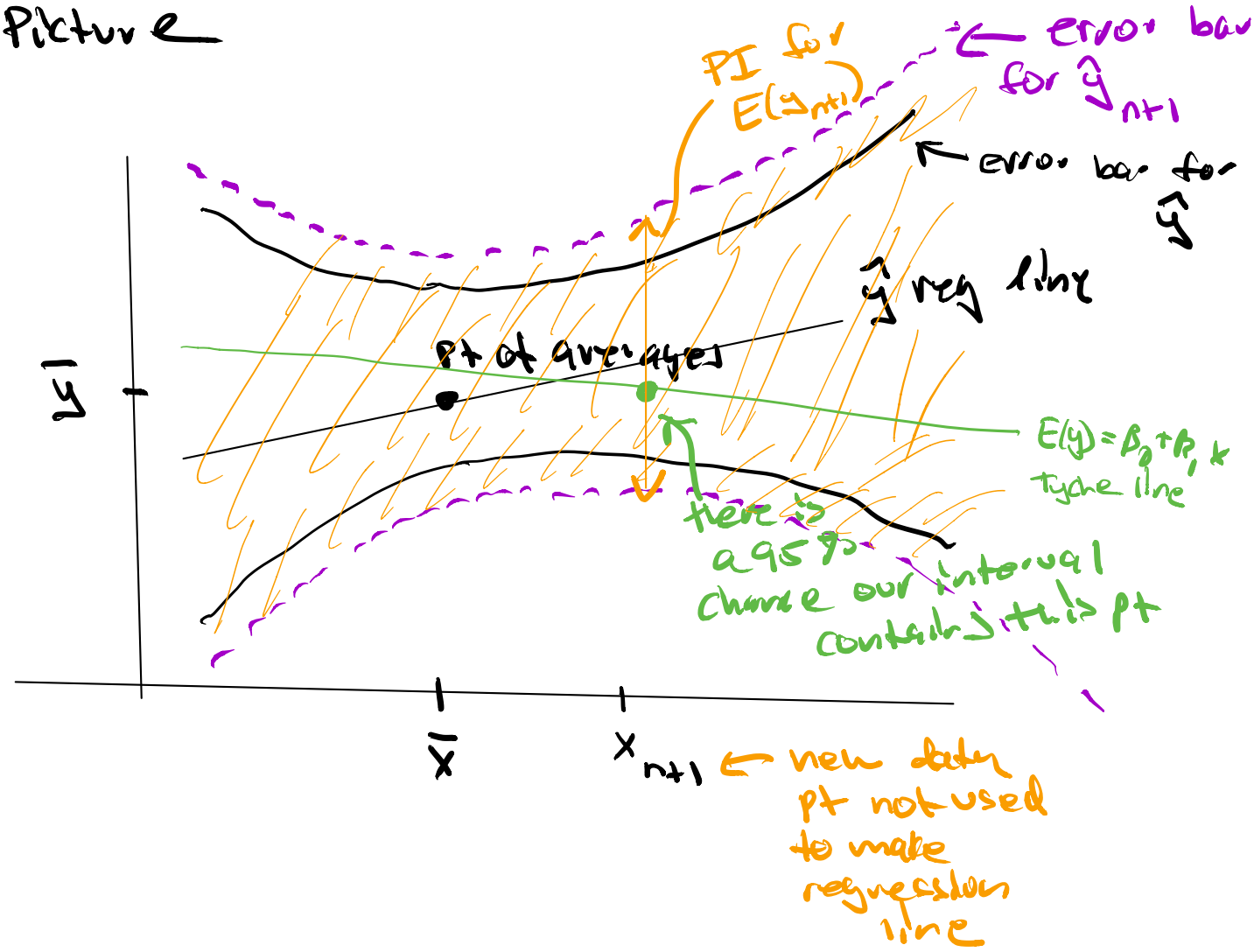
Thm A 95% PI for a new observation y_{n+1} for predictor x_{n+1} is

$$\hat{y}_{n+1} \pm t_{n-2}(0.025)\, s \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\,\mathrm{var}(x)}}$$

s: $\frac{RSS}{n-2}$

This is similar to 95% CI of $E(y)$ but has more uncertainty.

Picture

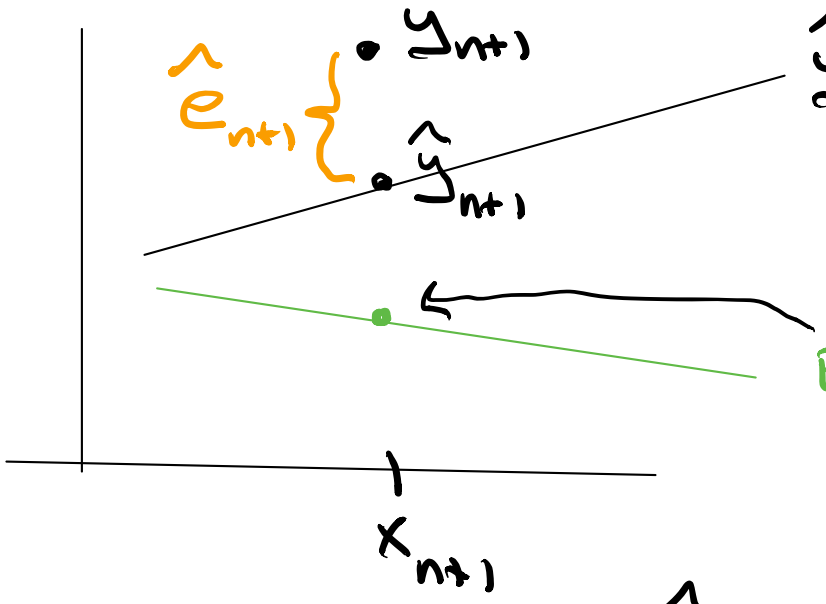


Notice that the error bar is wider for predicting $\hat{y}_{n+1}$ given $(x_1, y_1), \ldots, (x_n, y_n)$ and x_{n+1}.

Proof:

When y_{n+1} is a new response, not used in the calculation of the regression line, it is independent of $\hat{y}_{n+1}$. Consequently the variance of the residual $\hat{e}_{n+1}$ is larger than the variance of the residual $\hat{e}_i$ for $1 \le i \le n$.

Picture



$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ regression line from data $(x_1, y_1), \ldots, (x_n, y_n)$

$E(y_{n+1}) = \beta_0 + \beta_1 x_{n+1}$ Tyche line.

The variance of $\hat{e}_{n+1}$ is larger than the variance of $\hat{y}_{n+1}$.

We will see that

$$\operatorname{var}(\hat{e}_{n+1}) = \operatorname{var}(y_{n+1}) + \operatorname{var}(\hat{y}_{n+1})$$

$$= \sigma^2 + \frac{\sigma^2}{n} + \frac{(x_{n+1} - \bar{x})^2 \sigma^2}{n \operatorname{var}(x)}$$

$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n \operatorname{var}(x)}\right)$$

So why is y_{n+1} and $\hat{y}_{n+1}$ indep?

$\hat{y}_{n+1}$ is a pt on regression line that was calculated with $(x_1, y_1), \ldots, (x_n, y_n)$ not (x_{n+1}, y_{n+1}).

we have $E(\hat{e}_{n+1}) = E(y_{n+1} - \hat{y}_{n+1})$

$= E(y_{n+1}) - E(\hat{y}_{n+1}) = \boxed{0}$

$E(\hat{y}_{n+1}) = E(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1})$

$= \beta_0 + \beta_1 x_{n+1}$

$= E(y_{n+1})$

so $y_{n+1} - \hat{y}_{n+1} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\,var(x)}\right)\right)$

$$\Rightarrow \frac{y_{n+1} - \hat{y}_{n+1} - 0}{s\sqrt{\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\,var(x)}\right)}} \sim t_{n-2}$$

$$\Rightarrow \hat{y}_{n+1} \pm t_{n-2}^{(0.025)}\, s\sqrt{\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\,var(x)}\right)}$$

is a 95% PI

□

— — — — — — break — — — — — —

F statistic for regression — not in book.

For HW 6 you will show

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2$$

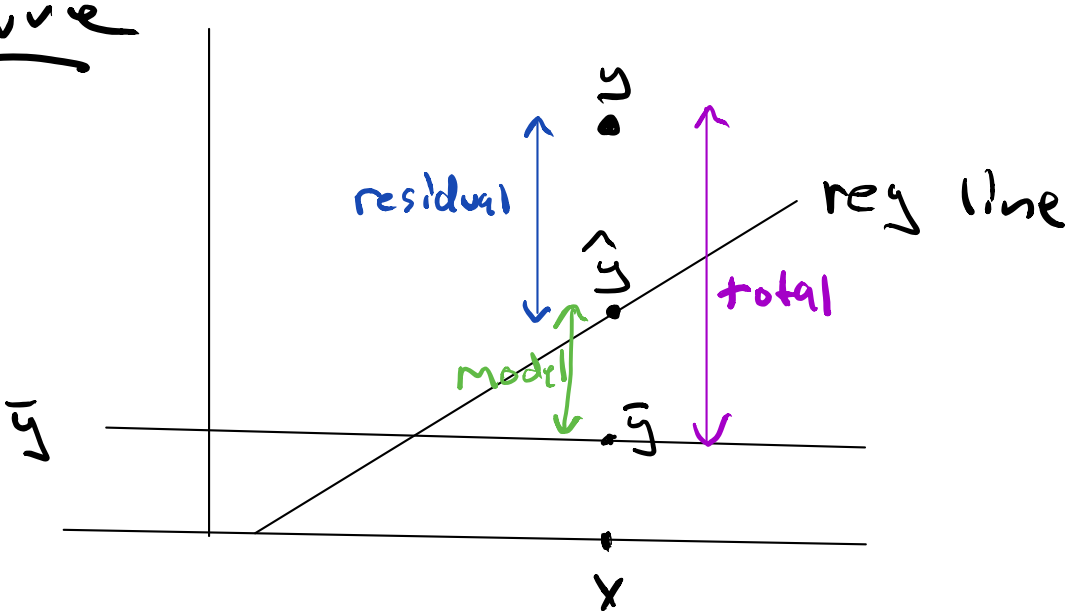
SS_{TOT} = TSS (total)

SS_{error} = RSS (residual)

SS_{model} = MSS (model)

i.e. we can divide the total variability of the dependent variable into two components: the variability of the regression line about the mean and the residual variation of observations about the regression line.

Picture



We showed RSS has $n-2$ d.f

TSS has $n-1$ d.f

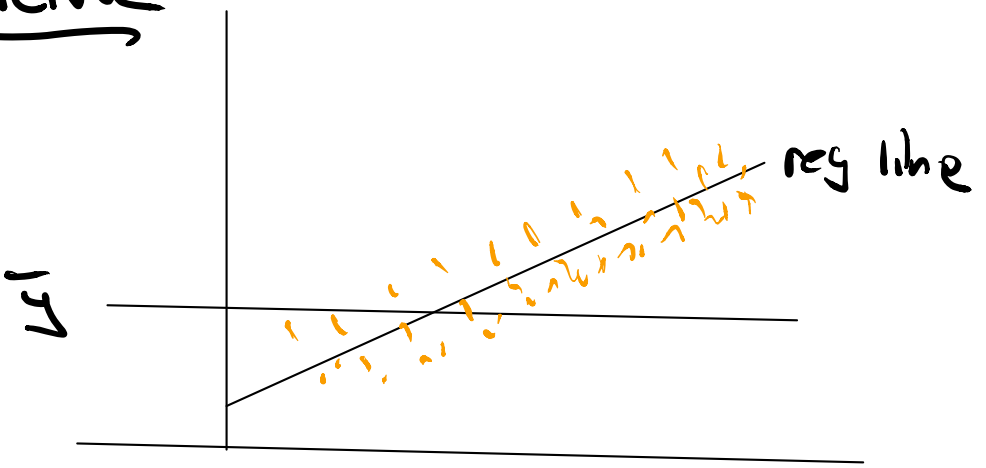
$\frac{RSS}{\sigma^2}$ and $\frac{MSS}{\sigma^2}$ are independent χ^2 so their d.f add. $\Rightarrow$ d.f MSS is 1.

This allows us to do hypothesis testing.

null $\beta_1 = 0$ (no relationship between x and y)

alt $\beta_1 \neq 0$

Picture



the spread of these $\hat{y}$ values about $\bar{y}$ is bigger the steeper β_1 gets

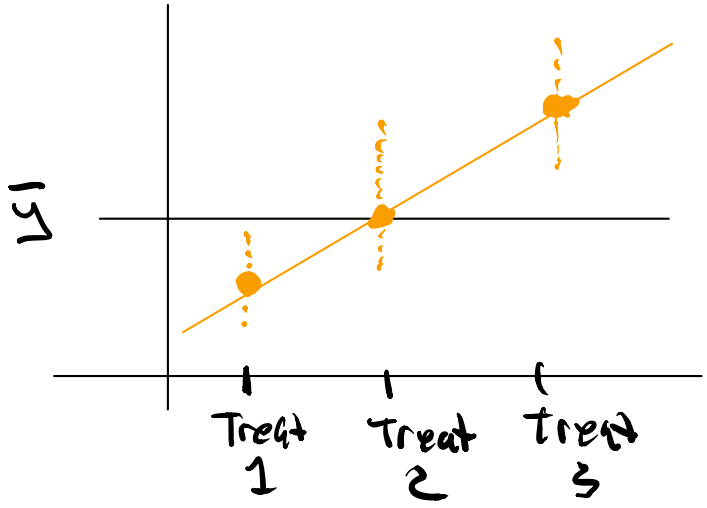
If the null is true ($\beta_1 = 0$) then MSS is very small.

For a fixed TSS if MSS is large then RSS is small.

An aside

Anova picture

$SST = SSW + SSB$



Informally, if you plot values of y for different treatment grps, the regression line goes through the pt of averages for each grp and you can see that
RSS = SSW
MSS = SSB.
so ANOVA really is regression.

End of aside,

Recall (p 194 Chap 6) def^n F dist:

Let U and V be indep χ^2 RV w/ d.f m and n respectively. Then

$$\frac{U/m}{V/n} \sim F_{m,n}$$

For us, with k predictors, k and $n-k-1$

$$\frac{\left(\frac{MSS}{\sigma^2}\right)/1}{\left(\frac{RSS}{\sigma^2}\right)/n-2} = \frac{MSS}{\left(\frac{RSS}{n-2}\right)} \sim F_{1,n-2}$$

we reject the null when

$F_{1,n-2}$ is large.

For simple linear regression F test is equivalent to t test

Not surprisingly

$$F_{1,n-2} = t^2_{n-2}$$ ← t-test for $\beta_1 = 0$.

R^2 – the coefficient of determination

In HW 6 you will show $\sigma^2_{\hat{e}} = (1-r^2)\sigma^2_y$

we have $\sigma^2_{\hat{e}} = \frac{RSS}{n} = \frac{\sum(y_i - \hat{y}_i)^2}{n}$

and $\sigma^2_y = \frac{TSS}{n} = \frac{\sum(y_i - \bar{y})^2}{n}$

so $RSS = (1-r^2)TSS$

$$\Rightarrow r^2 = 1 - \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS} = \frac{MSS}{TSS}$$

↑ called R^2 in multiple regression.

interpretation

R^2 is the fraction of the total variation of y "explained" by the model.

If MSS is very small the slope slope of the regression line is small, — i.e r^2 close to zero.

If MSS = TSS the data lies on a straight line, — i.e $r^2 = 1$

---------- End of simple linear regression ----------

Multiple linear regression is similar.

If we assume that our k predictors are independent we can do an F test whether $\beta_1, \beta_2, \ldots, \beta_k$ are all zero. For simple linear regression we only have β_1 and F test is equivalent to t test as mentioned above.

picture of the regression plane.

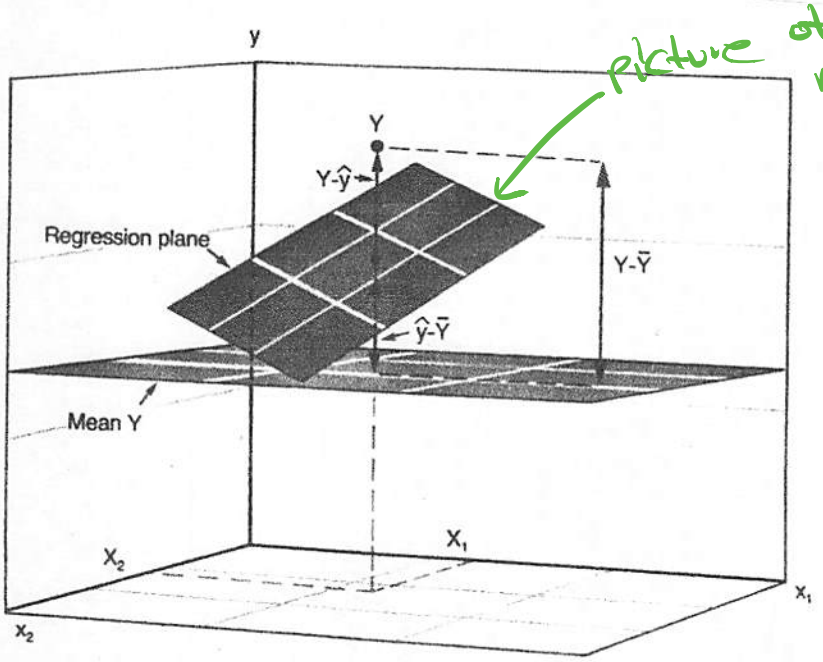


FIGURE 3-5 The deviation of the observed value of Y from the mean of all values of Y, $(Y - \bar{Y})$, can be separated into two components: the deviation of the observed value of Y from the value on the regression plane $(Y - \hat{y})$ at the associated values of the independent variables X_1 and X_2, and the deviation of the regression plane from the observed mean value of $\bar{Y}$ $(\hat{y} - \bar{Y})$ (compare with Fig. 2-7).

We have

$TSS = MSS + RSS$

$n-1$ d.f. $\uparrow$ K d.f. $\uparrow$ $n-K-1$ d.f.

just like we had before,

Let $R^2 = \frac{MSS}{TSS}$ coeff. of determination

Let $F = \frac{MSS/K}{RSS/n-K-1} = \frac{MSS/TSS \cdot K}{RSS/TSS \cdot (n-K-1)} = \frac{\frac{MSS}{TSS \cdot K}}{\frac{TSS - MSS}{TSS \cdot (n-K-1)}} = \boxed{\frac{R^2/K}{\frac{1-R^2}{n-K-1}}}$

this is our T.S.

If we reject the null it is natural to ask which of the β_i, $i=1,2,\ldots,k$ are statistically different from zero? In otherwords which of the variables $x_1, x_2, \ldots, x_k$ have a relationship with y.

You will explore all of this hopefully in your next stats class !!