§2.4 43] p31 $R^2 = [\text{cor}(Y, \hat{Y})]^2$ if a constant is in the model. $0 \le R^2 \le 1$. If $n \ge 5p$, high R^2 means the least squares "line" is much better at explaining $Y|\underline{x}$ than the line $Y = \bar{Y}$. (identity line or hyperplane)

44] p31 Suppose $x_1 \equiv 1$, the normal MLR model holds and $\beta_2 = \cdots = \beta_p = 0$. Then the line $Y = \bar{Y}$ should be used instead of the OLS "line" but $E\,R^2 = \frac{p-1}{n-1}$. So

if $n = 101$ and $p = 91$ then $E\,R^2 = .9$
51 $.5$
21 $.2$.

This is one reason $n \ge 5p$ is needed. (want R^2 near 0 if $\beta_2 = \cdots = \beta_p = 0$.)

45] know reject H_0 if pvalue $< \delta$. Use $\delta = 0.05$ if δ not given.

46] know for final The Anova F test tests whether the predictors $x_2, \ldots, x_p$ are needed in the model. If needed, use OLS. If not needed, use the line $\hat{Y} = \bar{Y}$. Assume a constant is in the model. The 4step test is

t test + CI know 43

i) H_0 $\beta_2 = \cdots = \beta_p = 0$ $\quad H_A$ not H_0

ii) $F_0 = \dfrac{MSR}{MSE}$

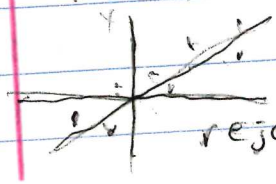
iii) $pval = P(F_{p-1,n-p} > F_0)$ } usually obtained from output

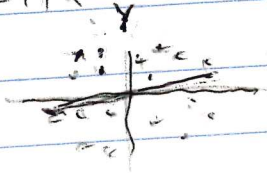
iv) Reject H_0 if $pval < \delta$. Then there is an MLR relationship between Y and the predictors $x_2, \ldots, x_p$.

Fail to reject H_0 if $pval \geq \delta$. Then there is not an MLR relationship between Y and $x_2, \ldots, x_p$.

Replace $Y, x_2, \ldots, x_p$ by names from story problem.

47] Response and residual plots should be good with $n \geq 5p$. If identity line fits data better than any horiz line, then the Anova F test will reject H_0. If the response plot looks like the residual plot so there is a horiz line that fits the data about as well as the identity line, then can't tell whether the Anova F test will reject H_0 or fail to reject H_0, but the MLR relationship is weak.





48) p32 know
Anova table

source	SS	df	MS	F	pval
model or regression	SSR	$p-1$	$MSR = SSR/(p-1)$	$F_0 = \frac{MSR}{MSE}$	for H_0 $\beta_2 = \dots = \beta_p = 0$
residual or error	SSE	$n-p$	$MSE = SSE/(n-p)$		
total	SSTO	$n-1$			

often omitted
See ex 2.5, 2.6

ex) mussel data $n=82$, $Y = \log(M)$, M = mussel muscle mass
$X_2 = \log(S)$ $X_3 = \log(H)$, S = shell mass
H = shell height

source	df	SS	MS	F	pval
reg	2	33.9208	16.9604	339.05	0.000
residual	79	3.9518	0.05063		

Perform the Anova F test

i) H_0 $\beta_2 = \beta_3 = 0$ H_A not H_0
ii) $F_0 = 339.05$
iii) pval $= 0.0$
iv) reject H_0, there is an MLR relationship between $\log(M)$ and the predictors $\log(S)$ and $\log(H)$

49) p30 If there is a constant in the MLR model, $SSTO = SSE + SSR$

$SSE = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ measures residual variability
(error (hat) sum of squares)

$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, $SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$

SSTO measures variability of the response Y; SSR: regression sum of squares

P32, 33, 36 Theorem 2.7 P_{H_0} = Prob, assuming H_0 is true

50] pvalue $= P_{H_0}$(obtaining a test statistic as extreme as the one observed, where "as extreme" depends on the form of H_A. (9.5)

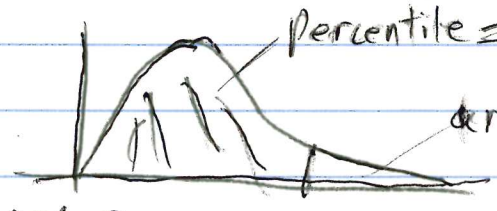
pval $= P(F_{p-1,n-p} > F_o)$ is given by the output.

pval = pvalue if the e_i are iid $N(0,\sigma^2)$, and if $n-p$ is large, then for many data sets, pval $\approx$ pvalue since $F_o \approx F_{p-1,n-p}$.

51] P32 $MSE = \hat{\sigma}^2 = \frac{SSE}{n-p}$ is an estimator of $\sigma^2 = V(e_i)$.

52] pvalue = 0 means it was impossible for the statistic to have occurred if H_0 is true, while pvalue = 1 means that it was impossible for the statistic to have occurred if H_A was true

53] F table

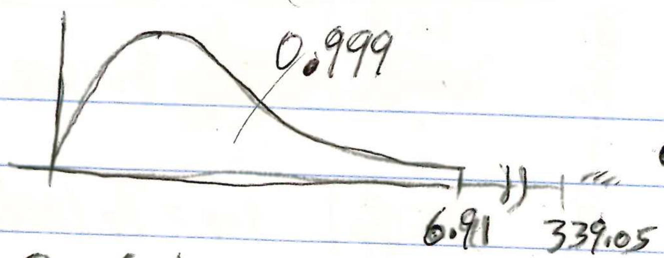


pval = 1 − table value

Know If $n-p > 60$, use den df = ∞, otherwise use closer df.

ex) mussels data $P(F_{2,79} > 339.05)$

SKIP until partial F test



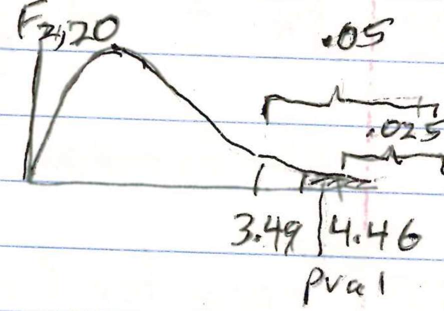
den	num 2
.999	6.91

pval ≈ 0 < 1 − .999 = .001

ex) Want pval $= P(F_{2,20} > 4.04)$

den df		num df 2
20	.95	3.49
	.975	4.46

} bracket test statistic



so .025 < pval < .05

ex) want pval $= P(F_{2,18} > 0.4)$

$\approx P(F_{2,20} > 0.4)$

		2
20	.5	.718

so pval > .5

54) For F table a) pval > 0.5 if Fo is smaller than the .50 value; b) pval is between 1 − big table val and 1 − small table value if Fo is between 2 table values
c) pval ≈ 0 < .001 if Fo value is larger than the .999 value

§2.7 p51, 63

55] (10.5

variable or label	estimate or coef	SE or Std err	t value	p value for H_0
intercept or constant	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$		$\beta_1 = 0$
x_2	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$t_{o2} = \hat{\beta}_2 / SE(\hat{\beta}_2)$	$\beta_2 = 0$
⋮	⋮	⋮		⋮
x_p	$\hat{\beta}_p$	$SE(\hat{\beta}_p)$	$t_{op} = \hat{\beta}_p / SE(\hat{\beta}_p)$	$\beta_p = 0$

56] p50 Know for final A $100(1-\delta)\%$ confidence interval (CI) for β_K is

$$\hat{\beta}_K \pm t_{n-p,\,1-\frac{\delta}{2}}\, SE(\hat{\beta}_K)$$

Use $\hat{\beta}_K \pm z_{1-\frac{\delta}{2}}\, SE(\hat{\beta}_K)$ if degrees of freedom $d = n-p > 30$.

57] p50 Know for final A 4 step Wald t test is

i) $H_0: \beta_K = 0$ $H_A: \beta_K \neq 0$

ii) $t_{oK} = \dfrac{\hat{\beta}_K}{SE(\hat{\beta}_K)}$

iii) $pval = 2P\left(t_{n-1} < -|t_{oK}|\right)$ } usually from output

iv) If $pval < \delta$, reject H_0. So x_K is needed in the MLR model for Y, given that the other predictors are in the model.

If $pval \geq \delta$, fail to reject H_0. So conclude x_K is not needed in the MLR model for Y, given that the other predictors are in the model.

Replace x_K by variable from story problem.