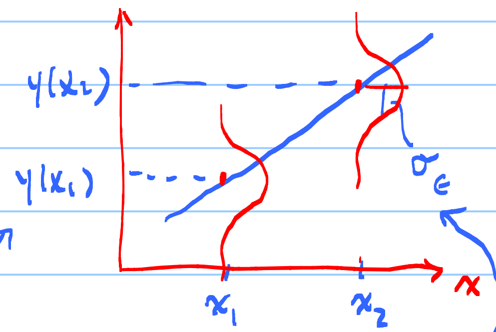
Lecture 25 (Ch. 11)

Last time: In preparation for doing inference in regression, we introduced The prob model for regression:

$$\boxed{\text{At a given } x,\ y \sim N(\mu, \sigma)}$$

$\mu = y(x)$, $\sigma = \sigma_\epsilon$



The μ param. (ie. center) of The Normal dist. of y's is allowed to vary with x.

The σ param. (denoted σ_ϵ) is not a function of x, and is estimated/approximated with $s_e = \sqrt{\frac{SSE}{n-(k+1)}}$ $k = \#$ of β's.

Then,

1) $y(x) = \alpha + \beta x + \cdots$ = true mean of y, at a given x.

2) $\hat{y}(x) = \hat{\alpha} + \hat{\beta} x + \cdots$ = estimated mean of y, given x

3) 95% of y's, at given x are within $y(x) \pm 1.96\,\sigma_\epsilon$

4) $\text{prob}(a<y<b) = \text{pr}\left(\frac{a-y(x)}{\sigma_\epsilon} < \underbrace{\frac{y-y(x)}{\sigma_\epsilon}}_{z} < \frac{b-y(x)}{\sigma_\epsilon} \right) = \cdots$ (Table I)

5) more (below)

e.g. $\hat{\beta}$ (and $\hat{\alpha}$) is now a random variable! like $\bar{x}$

It has a distribution! like $\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}})$

It has a prob! like $\text{pr}(\bar{x} > \bar{x}_{obs})$

⋮

We can build a CI for β (and α) like CI for μ_x

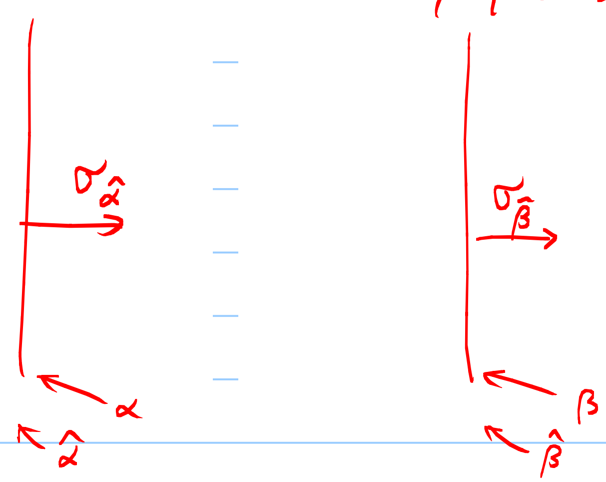
[redacted]

```
n = 10
n.trial = 64

x = c(1:n)
y_true = 10 + 2*x
sigma_eps = 15

par(mfrow=c(8,8),mar=c(0,0,0,0))
set.seed(123)
for(trial in 1:n.trial){
y_obs = y_true + rnorm(n,0,sigma_eps)
lm.1 = lm(y_obs ~ x)
plot(x, y_obs)
abline(10,2, col=2)
abline(lm.1, col=4)
}
```

Note that the x-values are the same across trials.
(in the kind of regression we are doing, x has no uncertainty; only y does.)



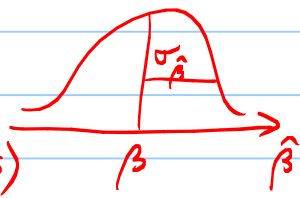
Let's build a CI (and hyp. test) for ONE β : $Y_i = \alpha + \beta x_i + \epsilon_i$

$\boxed{\text{Theorem:}}$ If $\epsilon \sim N(0, \sigma_\epsilon^2)$, Then $\hat{\beta}$ is normal with params:

Expected value (or mean) of The sampling dist. of $\hat{\beta}$

$$E[\hat{\beta}] \equiv \mu_{\hat{\beta}} = \beta \leftarrow \text{pop. slope}$$

$$\sqrt{V[\hat{\beta}]} \equiv \sigma_{\hat{\beta}} = \frac{\sigma_\epsilon}{\sqrt{S_{xx}}} \quad \text{(not obvious)}$$



$$\hookrightarrow S_{xx} = \sum_i^n (x_i - \bar{x})^2 = (n-1)\, s_x^2$$

Defn. of sample var.

$\boxed{\text{Ch. 7}}$

If $x \sim N(\mu_x, \sigma_x)$, Then $\bar{x}$ is Normal with params

$$E[\bar{x}] = \mu_{\bar{x}} = \mu_x$$

$$\sqrt{V[\bar{x}]} = \sigma_{\bar{x}} = \sigma_x / \sqrt{n}$$

Since $\hat{\beta} \sim N(\beta, \sigma_{\hat{\beta}})$, Then

$$z = \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}} = \frac{\hat{\beta} - \beta}{\sigma_\epsilon / \sqrt{S_{xx}}} \sim N(0,1)$$

$$t = \frac{\hat{\beta} - \beta}{s_e / \sqrt{S_{xx}}} \sim t\text{-dist.}$$

$df = n - \overset{k+1}{2}$

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t\text{-dist.}$$

$df = n - 1$

Then, from self-evident fact

$$pr(-t^* < t < t^*) = \text{Conf. level}$$

$df = n-2$ (Table VI)

C.I. for β : $\hat{\beta} \pm t^* \dfrac{s_e}{\sqrt{S_{xx}}}$

$H_0 : \beta = \beta_0$

$H_1 : \beta \ \square \ \beta_0$

$$t_{obs} = \frac{\hat{\beta}_{obs} - \beta_0}{s_e / \sqrt{S_{xx}}}$$

p-value $= (1,2) \cdot pr(\hat{\beta} \ \square \ \hat{\beta}_{obs}) =$

1 or 2-sided.

$= (1,2) \ pr(t \ \square \ t_{obs})$

$=$ Table VI, $df = n-2$

$$pr(-t^* < t < t^*) = \text{Conf. level}$$

$df = n-1$

C.I for μ : $\bar{x} \pm t^* \dfrac{s}{\sqrt{n}}$

$H_0 : \mu = \mu_0$

$H_1 : \mu \ \square \ \mu_0$

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

p-value $= (1,2)\ pr(\bar{x} \ \square \ \bar{x}_{obs})$

$= (1,2)\ pr(t \ \square \ t_{obs})$

$=$ Table VI, $df = n-1$

problem 11.17 [Revised]

$n=13$ $x=$ nickel content, $y=$ percentage austentite.

Data: $\Sigma(x_i-\bar{x})^2 = 1.183 = S_{xx}$

$\Sigma(y_i-\bar{y})^2 = 0.0508 = S_{yy} = SST$

$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 0.2073 = S_{xy}$

Question: Is There a statistically significant ($\alpha=0.05$) relationship between x and y? Hint: $SS_{exp} = \hat{\beta} S_{xy}$

1) C.I. β : $\hat{\beta} \pm t^* S_e/\sqrt{S_{xx}}$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{.2073}{1.183} = .1752$. $SSE = SST - SS_{exp}$

$= .0508 - (.1752)(.2073) = .014$

$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{.014}{13-2}} = 0.0357$

$\therefore$ 95% CI for β: $.1752 \pm \underset{df=13-2}{2.201}\left(\frac{.0357}{\sqrt{1.183}}\right) = 0.0328$ $= (0.10, 0.24)$

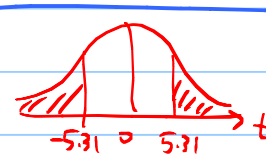
1) We are 95% Confident That The pop. β is in here.

2) There is a 95% prob That a random CI will cover β.

3) Corrollary: Relationship is statistically significant (zero not in CI).

2) $H_0: \beta = 0$ $\quad t_{obs} = \frac{.1752 - 0}{.0328} = 5.31$,

$H_1: \beta \neq 0$



p-value $= 2\,pr(\hat{\beta} > \hat{\beta}_{obs}) = 2\,pr(t > t_{obs})$

$= 2\,pr(t > 5.31) < 0.001$

p-value $< \alpha$ $\qquad$ Table VI df $= 13-2$

$\therefore$ Evidence That $\beta \neq 0$. (same conclusion as above).

In summary: We have 2 ways of testing whether there is a relationship between 2 continuous variables.

FYI

Note that the test of $\beta = 0$ is equivalent to testing if there is a linear relationship between x and y.
But if a linear relationship is all that you are testing, then we can test the population correlation coeff

$H_0: \rho = 0$

$H_1: \rho \neq 0$

The test statistic for this test is a bit weird:

$$\boxed{t = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}}}$$ has a t distr. with $df = n-2$.

Recall $r = S_{xy} / \sqrt{S_{xx} S_{yy}}$

This way, you take your data (x_i, y_i), compute the sample correl. coeff (r), then t_{obs}, and then p-value, all without any fitting.

3) For the above example:

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \cdots = .8456$$

$$t_{obs} = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}} = \cdots = 5.3$$ ← Same value as t_{obs} we got above when testing β.

p-value $= 2\,\text{prob}(t > t_{obs})$ = same as above.

$\therefore$ Same conclusion.

book problem

We have now done inference on β (and α),

What about multiple regression (ie. multiple x's and β's)?

In going from $y = \alpha + \beta x$ (1+1 parm)

to $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ (k+1 params) # of β's.

things generalize in a straightforward way.

Basically, all That happens is $df = n-2 \longrightarrow df = n-(k+1)$

This happens every where, e.g.

1) The estimate of σ_ϵ^2 is $s_e^2 = \dfrac{SSE}{n-(k+1)}$

2) The df associated with t-test changes: $n-2 \rightarrow n-(k+1)$

Finally, don't forget That The issues of collinearity, interaction, non-linearity, ..., and overfitting all return when doing multiple reg.

But, The presence of multiple β's allows for 1.5 more tests:

0.5) $\begin{cases} H_0: \beta_i \,\square\, \beta_0 \\ H_1: \beta_i \,\square\, \beta_0 \end{cases}$ Is The i^{th} predictor useful?

I'm saying 1.5 tests because this one is a straightforward generalization of The t-test for a single β (see next page).

1) $\left\{ \begin{array}{l} H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: \text{At least 1 } \beta_i \text{ is } \neq 0 \end{array} \right\}$ Are any of The predictors useful? (Test of "model utility".)

In $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k$,
if all $\beta_i = 0$, Then none of the predictors are useful for predicting Y.

Good News: The test for each β_i is The same as The t-test for a single β, except for $\boxed{df = n-(k+1)}$

E.g. suppose we want to test β_3:

$H_0: \beta_3 \square \beta_0$ (e.g. 0)

$H_1: \beta_3 \square \beta_0$

$$t_{obs} = \frac{\hat{\beta}_3 - \beta_0}{S_e/\sqrt{S_{x_3x_3}}}$$

p-value = (1,2) pr(t $\square$ t_{obs})

= --- df = n-(k+1)

Table VI

C.I. for β_3:

$$\sqrt{\frac{SSE}{n-(k+1)}}$$

$$\hat{\beta}_3 \pm t^* \frac{S_e}{\sqrt{S_{x_3x_3}}}$$

df = n-(k+1)

Technically, in multiple regression $S_{\hat{\beta}}$ is NOT $S_e/\sqrt{S_{xx}}$. The denominator ends-up being a more complicated function of x's. But when The predictors are completely uncorrelated, Then This formula is OK.

Note: even Though we are testing ONE β_i, The df is $n-(k+1)$

Bad News: If you test each of The β_i, separately, it's almost guaranteed That some of The β's will pass the test, ie. give small p-value, ie. are found to be useful, when infact, They are not.

Here is The proof, but it's only FYI

Bad News: If you test each of The β_i separately, you will make many more Type I errors Than α% of the time!

Consider 3 β's: β_1, β_2, β_3

Type I errors: $\underbrace{(\beta_1 \neq 0 \mid \beta_1 = 0)}_{e_1}$ $\underbrace{(\beta_2 \neq 0 \mid \beta_2 = 0)}_{e_2}$ $\underbrace{(\beta_3 \neq 0 \mid \beta_3 = 0)}_{e_3}$

You may commit The errors e_1 OR e_2 OR e_3 OR (e_1 and e_2) OR (e_1 and e_3) OR (e_2 and e_3) OR (e_1 and e_2 and e_3).

It can be shown That The prob of making at least 1 Type I error approaches 1 as The number of tests increases.

$\boxed{\text{Good News}}$: Enter The test of model utility!

Thm. $F = \dfrac{R^2/(k)}{(1-R^2)/(n-(k+1))} \sim$ F-distribution with $df = (k, n-(k+1))$

"numerator df"

"denominator df"

$\therefore$ p-value $= pr(F \geq F_{obs})$ Just like in 1-way ANOVA where H_1: Atleast ...

Then, if p-value $< \alpha$, we can reject H_0 $(\beta_1 = \beta_2 = \cdots = \beta_k = 0)$ in favor of H_1 (atleast 1 β_i is not zero).

This F-test allows you to do ONE test to find out if any of The predictors are useful for predicting y. This is very useful if k is large, because it tells you if any of The predictors are useful. I.e. it tells you if There is a "needle in The haystack," to begin with!

$\boxed{\text{IF}}$ you get a significant result (ie. p-value $< \alpha$) from The test of model utility, Then There is evidence That atleast one of The predictors is useful. $\boxed{\text{THEN}}$ you can do separate tests on each of The β's to see which predictors are useful. (see next page).

But $\boxed{\text{IF}}$ The F-test comes back as non-significant, Then There is no evidence That any of The predictors are useful. $\boxed{\text{THEN}}$, you don't have to test each predictor, separately. This will not only save time, but more importantly, it will save you from The danger of making multiple Type I errors (ie. declaring some predictor as useful, when infact, it is not).

Recall That

- "bad" Things happen if you keep adding terms to a regression model. Specifically, overfitting happens.
- overfitting is not a black and white Thing – it happens gradually, and in degrees, as you add more terms.
- even a complete "garbage" term can lead to overfitting.

What happens to F (and its p-value)?

more terms $\rightarrow$ higher $R^2 \rightarrow$ higher F $\rightarrow$ lower p-value.

I.e. If you keep Throwing enough predictors into a model (regression or otherwise), The F-test of model utility will find at least 1 useful predictor, regardless of whether or not The predictors are actually useful.

So, you MUST be Thoughtful about adding terms to regression
↳ ANY model!

FYI

The k-dependence of The formula for F does complicate Things a bit but you can ignore it, because The real problem arises from R^2 approaching 1, as The # of predictors increases.

Still, we can pay attention to The k-dependence:

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{R^2}{1-R^2}\left(\frac{n-(k+1)}{k}\right) = \frac{R^2}{1-R^2}\left(\frac{n-1}{k}-1\right)$$

Now, technically k must be less Than $(n-1)$, otherwise $F<0$, which it cannot be. So, $k<n-1$, in which case The largest allowed value of k is $n-2$, and so $\frac{n-1}{k}$ is at most $\frac{n-1}{n-2}$, ie, a constant! Then, we're back to looking at how R^2 grows.

11.66

```
The regression equation is
durpr = -0.912 + 0.161 formconc + 0.220 catratio + 0.0112 temp + 0.102 time

Predictor        Coef       StDev        T       p
Constant      -0.9122      0.8755    -1.04   0.307
formconc      0.16073     0.06617     2.43   0.023
catratio      0.21978     0.03406     6.45   0.000
temp         0.011226    0.004973     2.26   0.033
time          0.10197     0.05874     1.74   0.095
S = 0.8365 R-Sq = 69.2% R-Sq(adj) = 64.3%

Analysis of Variance
Source        DF      SS       MS       F      P
Regression     4 39.3769   9.8442  14.07  0.000
Error         25 17.4951   0.6998
Total         29 56.8720
```

You have already learned what all These numbers are, from The prelab. But now, we are going to do everything by hand.

Also $F = \frac{MS_{expl}}{MS_{Error}} = \frac{9.8442}{0.6998}$

$n-(k+1)$

$= n-1$

a) Is The model useful?

F-test: $F_{obs} = \frac{R^2/k}{(1-R^2)/(n-(k+1))} = \frac{\left(\frac{.692}{4}\right)}{\left(\frac{1-.692}{30-(4+1)}\right)} = 14.04$

p-value = prob$(F > F_{obs})$ = prob$(F > 14.04)$ $< .001$

According To Table VIII, df=(4,25)

∴ At any reasonable α, we can reject H_0 (That all $\beta_i = 0$) in favor of H_1 (That at least 1 of The $\beta_i \neq 0$). I.e. The model is useful.

b) Estimate, in a way that conveys info about precision & reliability, The average change in durability press rating associated with a 1-degree increase in curing temperature, when all other predictors remain fixed. (if There is NO collinearity)

I.e. what's The C.I for β_{temp} , t^* at $df = n-(k+1) = 25$

95% C.I.: $\hat{\beta} \pm t^* \frac{se}{\sqrt{S_{xx}}}$ = .0112 ± 2.060 $\frac{0.8365}{\sqrt{?}}$ ← not given!

2-sided

S_{xx} ← temp.

std. err in $\hat{\beta}$ from printout.

∴ .0112 ± 2.060 (.004973) ⟹ (.001, .021)

This is The interval estimate of β_{temp}. It's useful as it is, but we can also see That $\beta_{temp} \neq 0$

We can build The CI for all The other β_i:

C.I.	form conc	.1607	± 2.060	(.06617)	=	(0.02, 0.30)
	cat ratio	.2198	± "	.03406	=	(0.15, 0.29)
	temp	.0112	± "	.00497	=	(0.001, 0.02)
	time	.10197	± "	.0587	=	(−0.02, 0.22)

Note That 3 β's are non-zero.
At least 1

✓

Given That There is no evidence That "time" is a useful predictor, you may remove it from The regression model so That the "smaller" model will be less likely to overfit data.

In part a, we found out that atleast one of the $\beta_i \neq 0$.
To see which one(s), we test each of them!

$H_0: \beta_i = 0$ vs. $H_1: \beta_i \neq 0$ for each i.

c) E.g. $H_0: \beta_{formald.} = 0$

$H_1: \beta_{formald.} \neq 0$

$t_{obs} = \dfrac{.16073 - 0}{\boxed{.06617}} = 2.43$ (Check The output!)

$\dfrac{s_e}{\sqrt{S_{xx}}}$

formald.

(from Table)

$p\text{-value} = 2\, prob(t > t_{obs}) = 2(.012)$

even though testing 1 β.

$\boxed{df = n - (k+1) = 25}$

$= .024$ (check output!)

So, p-value $< \alpha$ $\Longrightarrow$ formaldehyde provides useful info.

In fact, look at all The p-values:

.023 , .000 , .033 , .095

look at last Col. of printout.

At $\alpha = .05$ $\underbrace{\beta_{formald.} \neq 0 \quad \beta_{cat.} \neq 0 \quad \beta_{temp} \neq 0}_{\text{consistent with The conclusions in part b.}}$ β_{time} = cannot Tell

---

FYI

Note These p-values are different from what you would get if you did $y = \alpha + \beta_1 x_1$, $y = \alpha + \beta_2 x_2$, ... Etc. and tested if each of These β_i are zero. The multiple regression model is more correct because it does take into account The correlations between predictors. See Ch. 3 lects.

hw_lect25_1
We have learned that if p-value < alpha, then there's evidence to reject H0 in favor of H1. For the test of model utility, p-value = pr(F > F_obs). So, for that p-value to be less than alpha, F_obs must be larger than some critical value.

a) At alpha=0.05, find the critical value of F_obs for a multiple regression problem involving four betas, and 30 cases.

b) Find the critical value of R^2 (above which p-value < alpha). Hint: The F-ratio appearing in the test of model utility depends on R^2 of the model. So, if you know the critical value of F (as in part a), then you know the critical value of R^2.

Moral: Like all other tests we have studied, the reject/no-reject decision can be based on the critical value of some statistic, i.e. without a p-value. For the test of model utility, the decision can be made by comparing F_obs with some critical value (e.g. found in part a), or even by comparing R^2_obs with its critical value (e.g. found in part b).

hw_optional
We have seen that adding useless predictors to a regression model will increase R2. Here, let's examine what our inference methods say if the predictors are, in fact, useless. Suppose the true/pop fit is y = 1,(i.e., no x at all), and so a possible sample from the population could be the following:

```
set.seed(123)     # Use this line to make sure we all get the same answes.
n = 20
y = 1 + rnorm(n,0,1)
```

a) Write code to make data on 10 useless predictors (and no useful predictors) each from unif(-1,+1), fit the model y = alpha + beta1 x1 + ... + beta10 x10, perform the test of model utility, and perform t-tests on each of the 10 coefficients to see if they are zero. Show/turn-in your R code.

b) According to the F-test of model utility, are any of the predictors useful at alpha = 0.1?

c) According to the t-tests, are any of the predictors useful at alpha = 0.1? See the solns to make sure you understand the moral of this exercise.

hw_optional
Consider a multiple regression problem with k betas on the right-hand side. Suppose all of the k predictors are completely useless. But, of course, we don't know that, so we test each of the betas individually. Our hyp. testing formalism assures that each test has prob. alpha of finding the predictor useful (when in fact it's useless).
a) what's the prob. of finding j useful predictors out of k predictors? Hint: Here you should recognize a familiar string of words here!
b) What's the prob. that at least 1 of the k predictors will be found to be useful (when it's not)?
c) Plot that prob vs. k = 1:100, for alpha=0.05, and for alpha=0.01
(Make sure you check the soln, later)