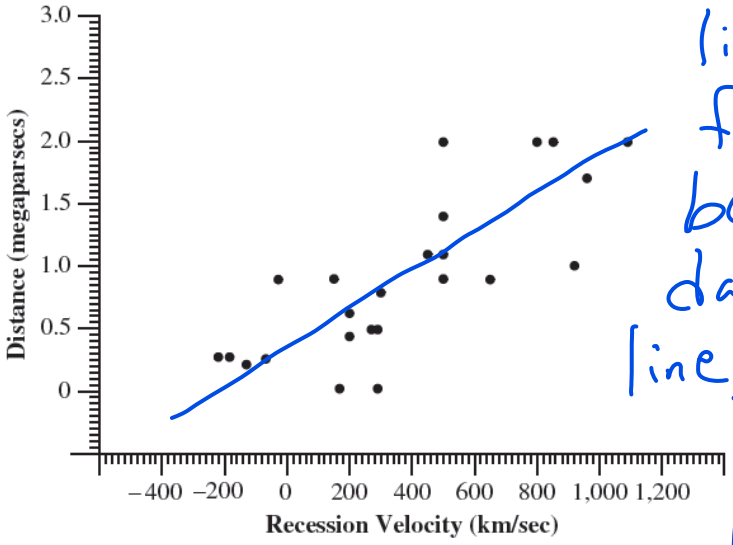
ST 411/511 Outline 7

Reading assignment: Chapter 7. This chapter introduces simple linear regression: making inference about a line fit to bivariate data. two values per obs.

Chapter 7 Simple Linear Regression: A Model for the Mean

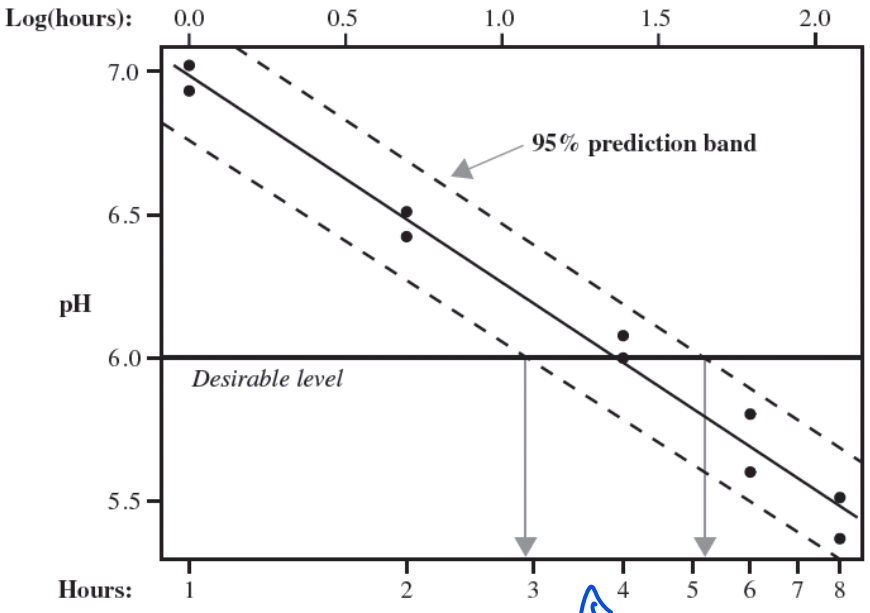
Case Study 7.1.1: Distance vs. velocity of extra-galactic nebulae.

DISPLAY 7.1 Scatterplot of measured distance versus velocity for 24 extra-galactic nebulae



linear regression finds line that best describes data, and use line to make inference about pop'n that gave the points.

DISPLAY 7.4 Meat processing data with estimated regression line (from the simple linear regression of pH on log time after slaughter) and a 95% prediction band



log scale

Structure of data in Chapters 1-6:

```
> head(case0601)
  Score Handicap
1   1.9      None
2   2.5      None
3   3.0      None
4   3.6      None
5   4.1      None
6   4.2      None
```

Score is response var., a numeric measurement on each subject.

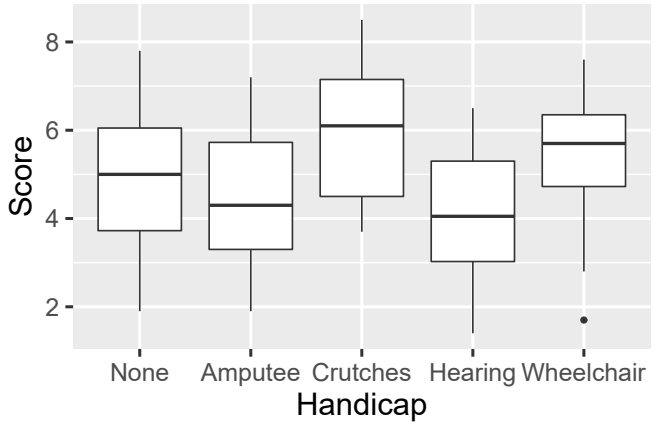
Handicap is a grouping var. a categorical explanatory var.

indep. var.

explanatory var.

response var

```
> qplot(Handicap,Score,data=case0601,geom="boxplot")
```



response is dependent var, dependent on explanatory var.

(this use of "dependence" has nothing to do with indep. assumption.)

```
> is.factor(case0601$Handicap)
[1] TRUE
> summary(case0601$Handicap)
      None    Amputee   Crutches    Hearing Wheelchair
        14         14         14         14         14
```

names and order of groups and sample sizes.

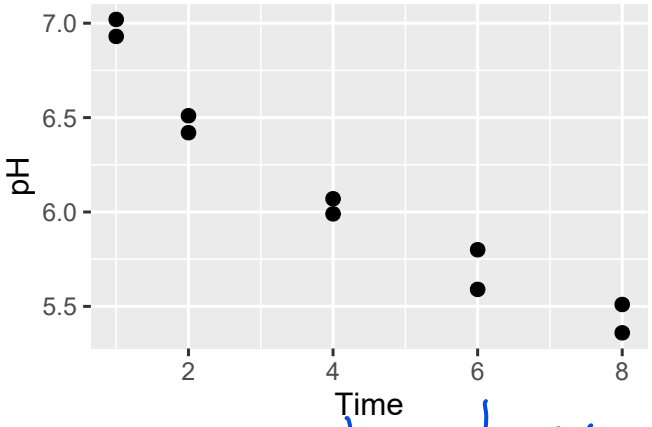
Structure of data in Chapters 7-8:

pH is response var.
Time is a numeric explanatory var.

```
> head(case0702)
  Time   pH
1    1 7.02
2    1 6.93
3    2 6.42
4    2 6.51
5    4 6.07
6    4 5.99
```

explanatory var.
response var.

```
> qplot(Time,pH,data=case0702)
```



resp. var.

Scatterplot, not boxplots

explanatory var.

```
> is.factor(case0702$Time)
[1] FALSE
> is.numeric(case0702$Time)
[1] TRUE
> summary(case0702$Time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     2.0     4.0     4.2     6.0     8.0
```

Summary stats.

Notation and terminology:

Response variable = Y } dependent var. by convention

Explanatory variable = X }

"Predictor var." Given a value for X, can predict Y (with some error).

Expl. var is the independent var.

(X_i, Y_i) = ith observation

By convention, expl. var. is first

n = sample size = # pairs (X_i, Y_i)

= # points on scatterplot.

just one expl. var.

Model for Simple Linear Regression (SLR)

Separate means model from Chapter 5:

$$\mu\{Y_{ij}\} = \mu_i,$$

Popn mean of Y_{ij} is μ_i (could be diff. for each group)

and Y_{ij} is normally distributed with standard deviation σ.

SLR Model:

$$\mu\{Y_i|X_i\} = \beta_0 + \beta_1 X_i,$$

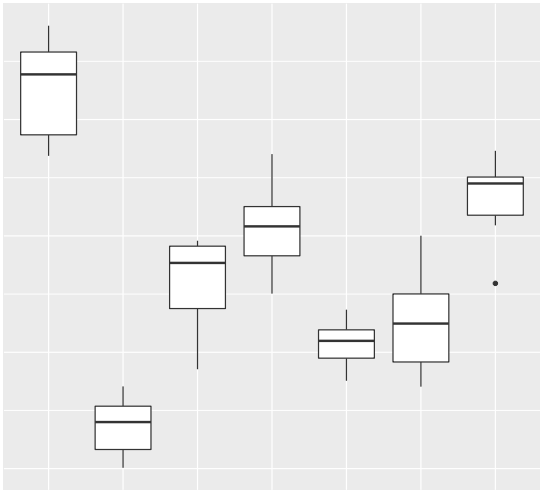
(where intercept line crosses vertical axis) slope

Popn mean of Y_i given X_i a linear function of X_i

and Y_i is normally distributed with standard deviation σ.

Three Models

Separate Means (ch. 5) $\mu\{Y_{ij}\} = \mu_i$ I = # groups



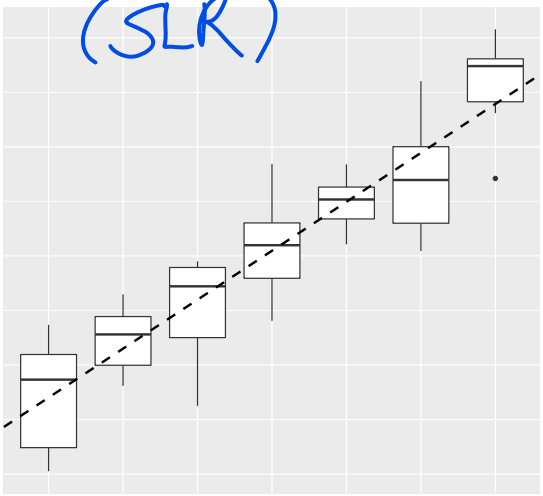
res. df = $n - I$

sample size − # mean parameters

Most complex (most mean parameters)

Simple Linear Regression (Ch. 7+8)

(SLR)



2 mean parameters

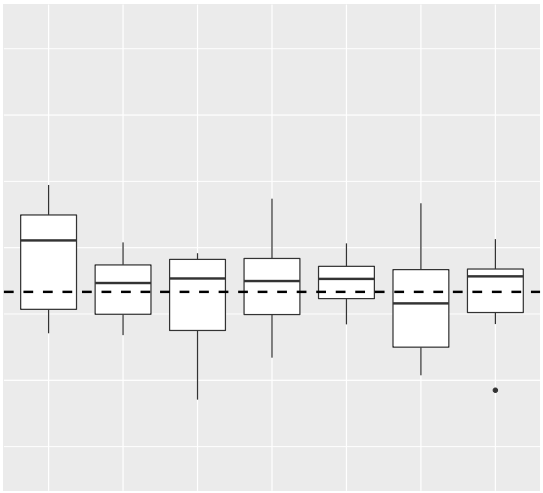
$\mu\{Y_i | X_i\} = \beta_0 + \beta_1 X_i$

Pop'n means fall on a line.

res. df = $n - 2$

Model of intermediate complexity. Means may differ, but fall on line.

Equal Means (ch. 5)

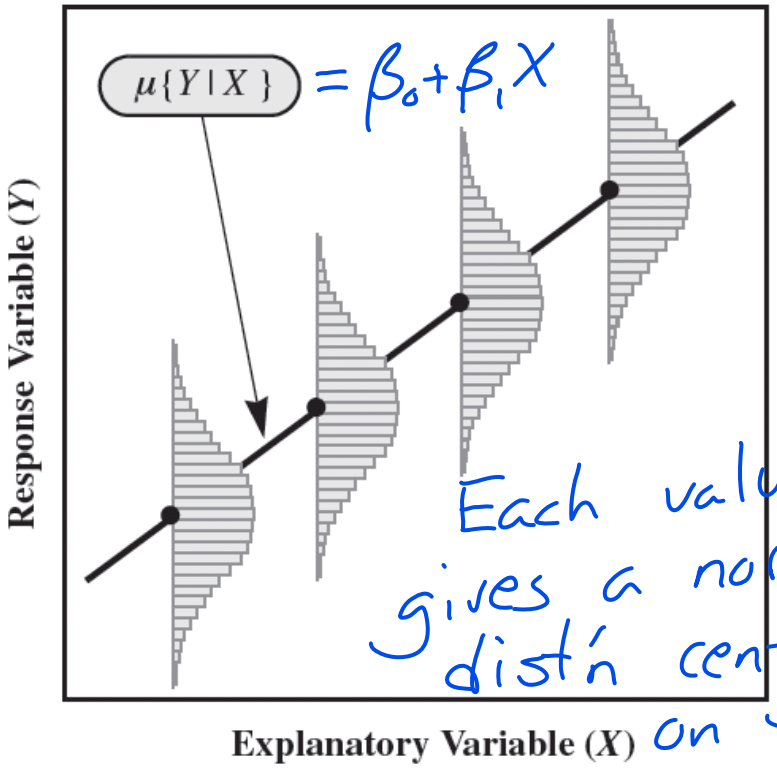


$\mu\{Y_{ij}\} = \mu$

res df. = $n - 1$

simplest model (only 1 mean parameter)

DISPLAY 7.5 The ideal normal, simple linear regression model



Each value of X gives a normal distn centered on the line.

Model Assumptions

1. There is a normally distributed subpopulation of responses for each value of the explanatory variable.
2. The means of the subpopulations fall on a straight line function of the explanatory variable.
3. The subpopulation standard deviations are all equal (to σ).
4. The selection of an observation from any of the subpopulations is independent of the selection of any other observation.

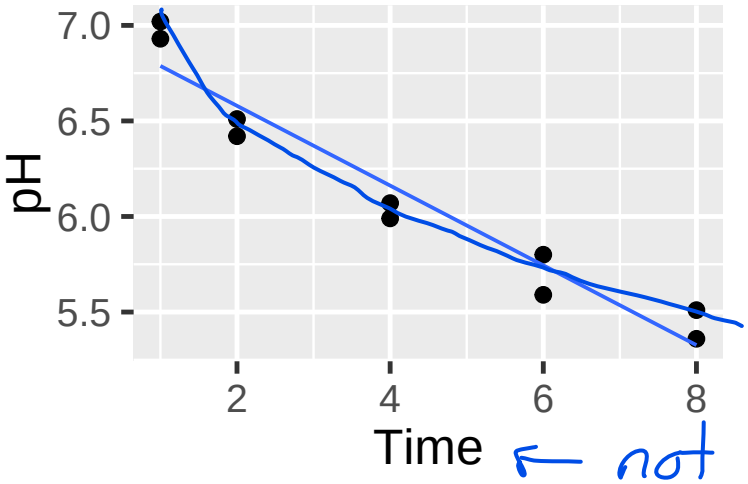
Same assumptions as in Ch. 5., except pop'n means fall on a line.

Aside: Y is random (as usual). X is fixed & not random.

See Lab 7 for this and more.

```
ggplot(case0702, aes(x=Time, y=pH)) +
  geom_point(size=I(4)) +
  geom_smooth(method=lm,se=FALSE)    # Add linear regression line
                                     # Don't add shaded confidence region
```

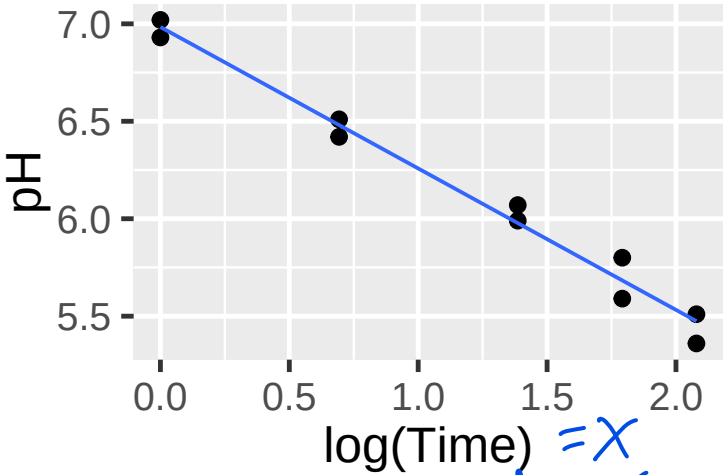
"linear model"



x-y relationship looks curved, not linear.

← not on log scale

```
ggplot(case0702, aes(x=log(Time), y=pH)) +
  geom_point(size=I(4)) +
  geom_smooth(method=lm,se=FALSE)    # Add linear regression line
                                     # Don't add shaded confidence region
```



Y = pH

This plot looks more linear. SLR looks more appropriate.

$X = \log(\text{Time})$

Model: $\mu\{pH \mid \log(Time)\} = \beta_0 + \beta_1 \log(Time)$

$\mu\{Y_i \mid X_i\} = \beta_0 + \beta_1 X_i$

Simple linear regression in R

Usual formula with response on left.

```
> case0702_lm <- lm(pH~log(Time), data=case0702)
```

lm object "linear model"

```
> summary(case0702_lm)
```

← asks for certain output

```
Call:
lm(formula = pH ~ log(Time), data = case0702)
```

← echos command

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.11466 -0.05888  0.02085  0.03612  0.11658
```

← can be used to assess assumptions. (Ch. 8)

β_0 & β_1 ↓ pt. ests. SE(pt.est.)

2-sided of t-tests
$H_0: \beta_0 = 0$
$H_0: \beta_1 = 0$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.98363    0.04853  143.90 6.08e-15 ***
log(Time)   -0.72566    0.03443  -21.08 2.70e-08 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

β_0
β_1

$s_p = \hat{\sigma}$

res. df = n − 2 = sample size − # mean parameters

```
Residual standard error: 0.08226 on 8 degrees of freedom
Multiple R-squared:  0.9823,Adjusted R-squared:  0.9801
F-statistic: 444.3 on 1 and 8 DF,  p-value: 2.695e-08
```

Multiple $R^2 = 0.9823$ proportion of variation in resp. explained by SLR model.

R^2 is always between 0 & 1. Close to 1 says pts are close to reg. line.

Adjusted $R^2 = 0.9801$ Like R^2 but takes into account model complexity.

F-test Compares SLR model (full model) to the equal means model (reduced model).

Unpacking the summary.lm output

Mean parameters: β_0 & β_1

Given x_i and these parameters, we know $\mu\{Y_i \mid x_i\} = \beta_0 + \beta_1 x_i$

Point estimates of mean parameters: (from coeff. table)

$\hat{\beta}_0 = 6.98363$

$\hat{\beta}_1 = -0.72566$

From R, not by hand.

Standard errors of the point estimates:

$SE(\hat{\beta}_0) = 0.04853$

$SE(\hat{\beta}_1) = 0.03443$

Estimated regression line:

$\hat{\mu}\{pH \mid \log(Time)\} = 6.98363 - 0.72566 \cdot \log(Time)$

SE's $\rightarrow$ (0.04853) (0.03443)

or $\hat{\mu}\{Y_i \mid X_i\} = 6.98363 - 0.72566\, X_i$ $\quad X_i = \log(Time)$

(0.04853) (0.03443) $\quad Y_i = pH$

Estimate of σ: $s_p = \hat{\sigma} =$ res. std error $= 0.08226$

(used to calculate $SE(\hat{\beta}_0)$ & $SE(\hat{\beta}_1)$, but we won't do this by hand.)

$\alpha = 0.01$

99% confidence interval for β_0: ← intercept parameter

$$\text{pt est} \pm t_{\text{df}}(1-\alpha/2)\ \text{SE(pt est)}$$

6.98363 ↙ ↘ 0.04853

res. df ↙

```
> qt(1-.01/2, 8)
[1] 3.355387
```

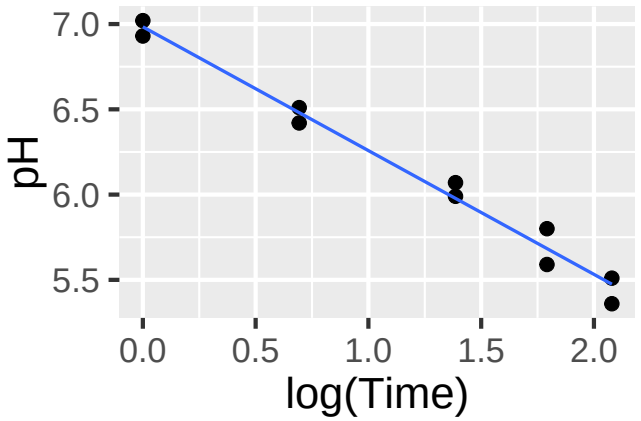
$6.98363 \pm 3.355387 \cdot 0.04853$

$\approx (6.82, 7.15)$

Statistical conclusion: A statistical conclusion for a confidence interval should contain three elements.

What is β_0 in our model?

- A statement indicating "estimation" and what quantity was estimated
- The endpoints of the confidence interval (point estimate optional)
- The confidence level



Intercept β_0 is point on line associated with $X = 0$

$\log(\text{Time}) = 0$

$\text{Time} = e^0 = 1$

We estimate the pop'n mean pH after 1 hour is 6.82 to 7.15 units (99% CI, SLR).

Some inferences (details later)

- Estimate β_0 and/or β_1. We just estimated β_0, the mean response when $X=0$. Could estimate β_1, which characterizes the relationship between pop'n mean Y and X.

- Estimate $\mu\{Y|X = X_0\}$. ← CI For any X_0, can estimate the pop'n mean of associated Y's. Pt. est. is the point on estimated regression line. Beware extrapolation – picking an X_0 far from what we have observed.

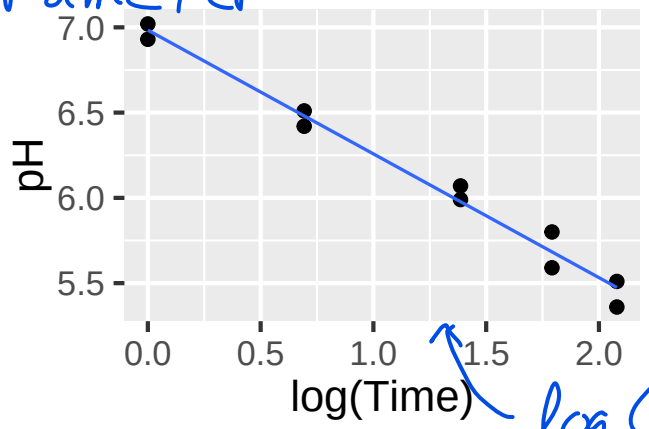
- Predict a new Y for a given value X_0 of X. ↑ used differently than "estimate". Given X_0, find a range of values that we would expect a random Y to fall in.

- Estimate or predict the X that yields a given value Y_0 of Y ("regression calibration"). Pick a pt and estimate (predict) associated time.

Estimating $\mu\{Y|X = X_0\}$. = popn mean Y for $X = X_0$

estimating the parameter

$$\text{pt est} \pm t_{\text{df}}(1-\alpha/2)\text{SE}(\text{pt est})$$ ← familiar formula



Can pick any X_0, but avoid extreme extrapolation.

$\log(4) = X_0 \approx 1.4$

Point estimate $\hat{\mu}\{Y|X = X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$

$\hat{\mu}\{Y|X = \log(4)\} \approx 6.984 - 0.726 \cdot \log(4) \approx 5.98$

$\overline{X}$ = average of obs. X's.

Sp ↓

Standard error $\text{SE}(\text{pt est}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$ ← scaling factor

We won't calculate by hand. ↑ sample size

What happens to SE(pt est) if

- $\hat{\sigma}$ is large? SE large. Makes sense: large $\hat{\sigma}$ means noisier data mean less certainty about pt. est.
- n is large? SE small. More data = more certainty.
- $(X_0 - \overline{X})^2$ is large? SE large. Less certainty if X_0 is far from center of X's.

DISPLAY 7.10 95% confidence interval for the mean pH of steers 4 hours after slaughter (from the estimated regression of pH on log(time) after slaughter for the meat processing data)

$$\hat{\mu}\{Y \mid 1.386\} = 6.9836 - 0.7257 \times 1.386 = 5.98$$

$$\text{SE}[\hat{\mu}\{Y \mid 1.386\}] = 0.08226 \sqrt{\frac{1}{10} + \frac{(1.386 - 1.190)^2}{9\,(0.6344)}} = 0.0269$$

$\hat{\sigma}$; $X_0 = \log(4)$; $\overline{X}$; n; $(n-1)$; s_X^2; $t_8(0.975)$

Upper limit: $5.98 + 2.306 \times 0.0269 = 6.04$

Lower limit: $5.98 - 2.306 \times 0.0269 = 5.92$

lm object

data frame containing value of Time

give SE(pt. est.)

```
> predict(case0702_lm, data.frame(Time=4), se.fit=TRUE,
+         interval="confidence")
$fit
       fit      lwr      upr
1 5.977651 5.915677 6.039625

$se.fit
[1] 0.02687513

$df
[1] 8

$residual.scale
[1] 0.08225969
```

pt. est

CI

Calculate a CI

lm object knows about logging Time.

= SE(pt. est.)

res. df = n − # mean parameters = n − 2 for SLR

$\hat{\sigma} = s_p$

Statistical conclusion:

We estimate popn mean pH of steers after 4 hours is 5.92 to 6.04 (95% CI, SLR).

("pH" implies units, so don't need them here.)

Comments

- We wrote a confidence interval for the mean response at a given value of the explanatory variable.

This is a CI for a pt. on pop'n regression line. CI quantifies uncertainty about est. line at that pt.

- pH is on a log scale.

This was original scale, so don't need to back-transform.

- Time was log-transformed.

Conclusion was in terms of original units (hours).

- $R^2 = 0.9823$ Very close to 1 (max possible R^2), so data fall close to estimated line. Not a lot of noise in data.

Not "estimating" because Y is random, not a parameter like a popn mean.

Predicting a future Y for a given value X_0 of X

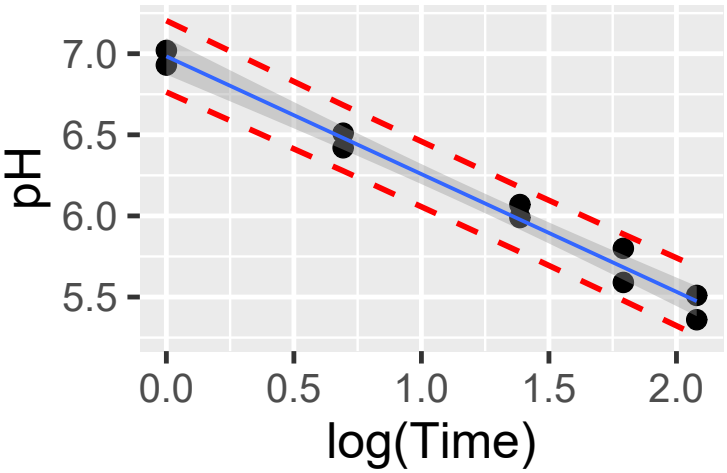
unobserved response

$$\text{pt pred} \pm t_{\text{df}}(1-\alpha/2)\text{SE}(\text{pt pred})$$

familiar form, but language is diff.

Prediction interval vs. confidence interval:

	Confidence Interval CI	Prediction Interval PI
Range of Plausible Values for	$\mu\{Y \mid X=X_0\}$ = pt. on popn. reg. line.	Y = random draw from normal distn centered at $\mu\{Y \mid X=X_0\}$
Uncertainty	Uncertainty about popn reg. line	Uncertainty about reg. line and where Y will fall in the normal distn.



Shaded band represents CI's "confidence band"

Dotted lines represent PI's "prediction band"

Point prediction $\hat{\mu}\{Y|X = X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$

Same as pt. est. for a CI of $\mu\{Y \mid X = X_0\}$. This is the pt. on est. line at $X = X_0$.

Standard error of the point prediction

Same as for SE(pt. est.)

$$\text{SE(pred)} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

$$[\text{SE(pred)}]^2 = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

$$= \hat{\sigma}^2 + \hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

$\hat{\sigma}^2$: est. of variance of normal distn. — Uncertainty about where in normal distn the new Y falls.

$\left[\text{SE(pt. est)}\right]^2$: Uncertainty in est. reg. line

DISPLAY 7.12 95% prediction interval for the pH of a steer carcass 4 hours after slaughter (from the estimated regression of pH on log time after slaughter)

$$\text{Pred}\{Y \mid 1.386\} = 6.9836 - 0.7257 \times 1.386 = 5.98 \quad (X_0 = \log(4))$$

$$\text{SE}[\text{Pred}\{Y \mid 1.386\}] = 0.08226 \sqrt{1 + \frac{1}{10} + \frac{(1.386 - 1.190)^2}{9\,(0.6344)}} = 0.0865$$

($\hat{\sigma}$; n; $\bar{X}$; $(n-1)$; s_X^2)

$t_8(0.975)$

Upper limit: $5.98 + 2.306 \times 0.0865 = 6.18$
Lower limit: $5.98 - 2.306 \times 0.0865 = 5.78$

Same as for CI except

```
> predict(case0702_lm, data.frame(Time=4), se.fit=TRUE,
+         interval="prediction")
$fit
       fit      lwr      upr
1 5.977651 5.778092 6.177209

$se.fit
[1] 0.02687513

$df
[1] 8

$residual.scale
[1] 0.08225969
```

PI

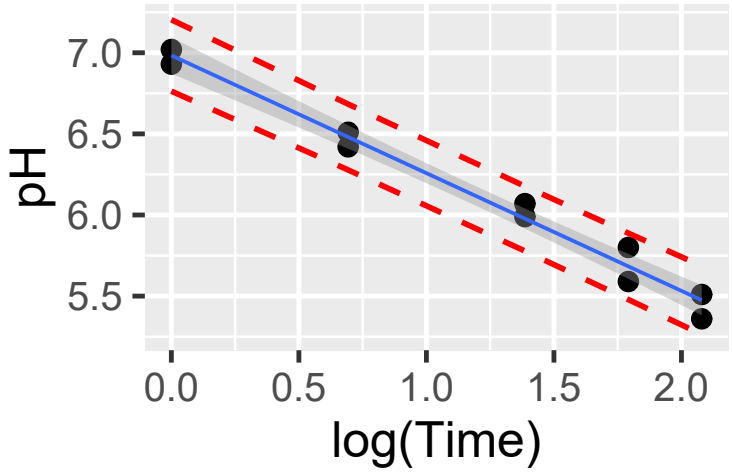
pt pred.

← Not SE(pred)
It's SE(pt. est.)

Statistical conclusion: It is predicted that the pH of a steer carcass after 4 hours is 5.78 to 6.18 (95% PI, SLR).

Note: "predicted" not "estimated"
No mention of pop'n mean.

Calibration: Estimating/predicting the X that yields a given value Y_0 of Y



Skip this.

$$\widehat{\mu}\{Y|\widehat{X}\} = \hat{\beta}_0 + \hat{\beta}_1\widehat{X}$$

The *Sleuth*'s approximate SEs

Calibration confidence interval:

$$\widehat{X} \pm t_{\text{df}}(1-\alpha/2) \cdot \frac{\text{SE}(\hat{\mu}\{Y|\widehat{X}\})}{|\hat{\beta}_1|}$$

For calibration prediction interval:

$$\widehat{X} \pm t_{\text{df}}(1-\alpha/2) \cdot \frac{\text{SE}(\text{pred}\{Y|\widehat{X}\})}{|\hat{\beta}_1|}$$

Calibration in R

```
> library(investr)
> calibrate(case0702_lm, y0=6, mean.response=TRUE)
estimate    lower    upper
1.355496 1.272344 1.442654
```

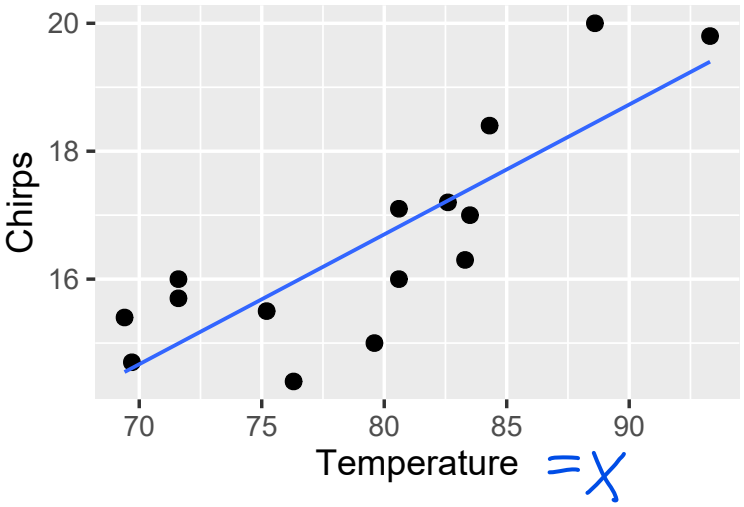
Statistical conclusion:

```
> calibrate(case0702_lm, y0=6)
estimate    lower    upper
1.355496 1.081071 1.633926
```

Statistical conclusion:

Interpreting β_1 = Slope parameter

Example: Number of cricket chirps per minute as a function of temperature.



Y = Chirps

X = Temperature

$\mu\{Y|X\} = \beta_0 + \beta_1 X$

β_1 quantifies relationship between X and $\mu\{Y|X\}$

What happens to the mean of Y if X is increased by one unit?

$$\mu\{Y|X+1\} = \beta_0 + \beta_1(X+1)$$

$$= \underbrace{\beta_0 + \beta_1 X}_{\mu\{Y|X\}} + \beta_1$$

So β_1 is change in mean response when explanatory var. increases by one unit.

```
> crickets.lm <- lm(Chirps~Temperature,data=crickets)
> summary(crickets.lm)

Call:
lm(formula = Chirps ~ Temperature, data = crickets)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6181 -0.6154  0.0916  0.7669  1.5549

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45931    2.98920   0.154 0.880239
Temperature  0.20300    0.03754   5.408 0.000119 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.986 on 13 degrees of freedom
Multiple R-squared:  0.6923,Adjusted R-squared:  0.6686
F-statistic: 29.25 on 1 and 13 DF,  p-value: 0.0001195
```

β_0 β_1 β_1 $SE(\beta_1)$ res. df

90% confidence interval for β_1: pt est $\pm\, t_{\text{df}}(1-\alpha/2)$ SE(pt est)

$\alpha = 0.1$

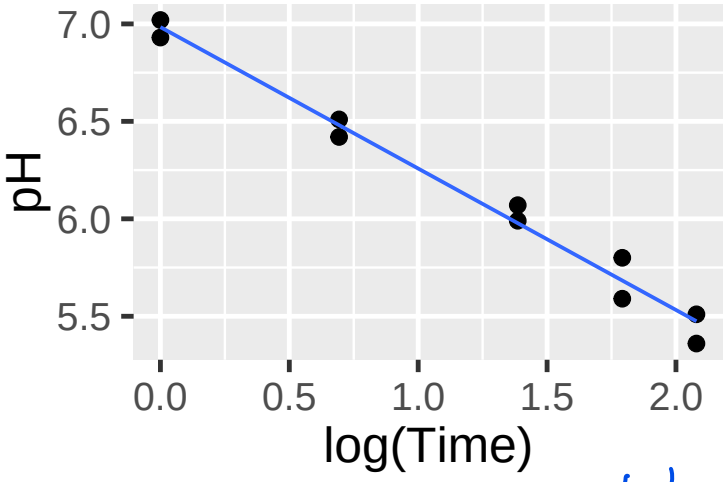
```
> qt(1-0.1/2,13)
[1] 1.770933
```

res. df

$0.203 \pm 1.770933 \cdot 0.03754 \approx (0.14, 0.27)$

Statistical conclusion: We estimate that for each 1-degree F increase in temp., pop'n mean # chirps per minute increases by 0.14 to 0.27 chirps. (90% CI, SLR).

Interpreting β_1 when predictor was log-transformed



β_1 is change in pop'n mean pH when log(Time) is increased by 1 unit. Not appropriate for stat. conclusion. Need to report in original time units.

```
> confint(case0702.lm,level=0.99)
                 0.5 %     99.5 %
(Intercept)  6.8207825  7.1464695
log(Time)   -0.8411714 -0.6101441
```

← 99% for β_1

Model: $\mu\{\text{pH}|\log(\text{Time})\} = \beta_0 + \beta_1 \log(\text{Time})$

log scale is multiplicative. Change in time should be multiplicative. Try doubling Time.

$$\mu\{pH \mid \log(2\cdot Time)\} = \beta_0 + \beta_1 \log(2\cdot Time)$$

$$= \beta_0 + \beta_1 [\log(2) + \log(Time)]$$

$$= \underbrace{\beta_0 + \beta_1 \log(Time)}_{\mu\{pH \mid \log(Time)\}} + \underbrace{\beta_1 \cdot \log(2)}_{\text{change in pop'n mean pH when Time doubled}}$$

```
> c(-0.8411714, -0.6101441) * log(2)
[1] -0.5830556 -0.4229197
```

← 99% CI for $\beta_1 \cdot \log(2)$

Statistical conclusion:

It is estimated that for each doubling of time, popn mean pH decreases 0.423 to 0.583 pH units (99% CI, SLR).

To interpret slope parameter, need to consider a meaningful change in explanatory variable.

When X logged, change is multiplicative.

When X not logged, change is additive.