Chapters 1 and 2

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Toluca data (p. 19)

- Toluca makes replacement parts for refrigerators.
- We consider one particular part, manufactured in varying lot sizes.
- It takes time to set up production regardless of lot size; this time plus machining & assembly makes up work hours.
- We want to relate work hours to lot size.
- $n = 25$ pairs (X_i, Y_i) were obtained.

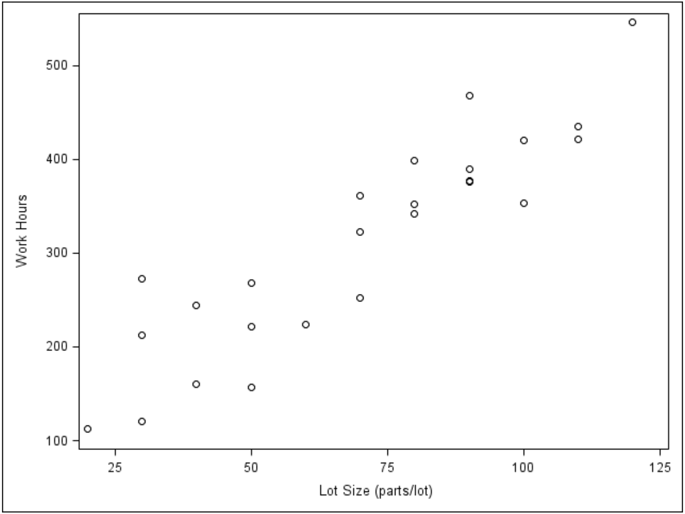
Toluca data, scatterplot & regression in SAS

```
data toluca;
input size hours @@;
label size="Lot Size (parts/lot)"; label hours="Work Hours";
datalines;
   80  399   30  121   50  221   90  376   70  361   60  224  120  546
   80  352  100  353   50  157   40  160   70  252   90  389   20  113
  110  435  100  420   30  212   50  268   90  377  110  421   30  273
   90  468   40  244   80  342   70  323
;
proc sgplot data=toluca;
scatter x=size y=hours; run;
options nocenter;
proc reg data=toluca; model hours=size; run;
```

Toluca data, SAS output

```
The REG Procedure
Dependent Variable: hours Work Hours

Number of Observations Read          25
Number of Observations Used          25

                         Analysis of Variance

                                   Sum of           Mean
Source                   DF       Squares         Square    F Value    Pr > F
Model                     1        252378         252378     105.88    <.0001
Error                    23         54825     2383.71562
Corrected Total          24        307203

Root MSE             48.82331    R-Square     0.8215
Dependent Mean      312.28000    Adj R-Sq     0.8138
Coeff Var            15.63447
                               Parameter Estimates

                                             Parameter       Standard
Variable     Label                   DF       Estimate          Error    t Value    Pr > |t|
Intercept    Intercept                1       62.36586       26.17743       2.38      0.0259
size         Lot Size (parts/lot)     1        3.57020        0.34697      10.29      <.0001
```

Toluca data



Roughly linear trend, no obvious outliers.

Toluca

The fitted model is

$$\widehat{\text{hours}} = 62.37 + 3.570 \times \text{lot size}.$$

- A lot size of $X = 65$ takes $\hat{Y} = 62.37 + 3.570 \times 65 = 294$ hours to finish, *on average*.
- For each unit increase in lot size, the mean time to finish increases by 3.57 hours.
- Increasing the lot size by 10 parts increases the time by 35.7 hours, about a week.
- $b_0 = 62.37$ is only interpretable for lots of size zero. What does that mean here?

Residuals & fitted values, Section 1.6

- The ith **fitted value** is $\hat{Y}_i = b_0 + b_1 X_i$.
- The points $(X_1, \hat{Y}_1), \ldots, (X_n, \hat{Y}_n)$ fall on the line $y = b_0 + b_1 x$, the points $(X_1, Y_1), \ldots, (X_n, Y_n)$ do not.
- The ith **residual** is

  $$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i), \quad i = 1, \ldots, n,$$

  the difference between observed and fitted values.
- e_i "estimates" ϵ_i.

Properties of the residuals (pp. 23–24)

1. $\sum_{i=1}^{n} e_i = 0$ (from normal equations)
2. $\sum_{i=1}^{n} X_i e_i = 0$ (from normal equations)
3. $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$ (1 and 2)
4. Least squares line always goes through $(\bar{X}, \bar{Y})$.

Estimating σ^2, Section 1.7

σ^2 is the error variance. A natural starting point for an estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2$. However,

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i - b_0 - b_1 X_i)^2 \\ &= \text{...a lot of hideous algebra later...} \\ &= \frac{n-2}{n}\sigma^2. \end{aligned}$$

So in the end we use the unbiased *mean squared error*

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2.$$

MSE and SSE

So an estimate of $\text{var}(Y_i) = \sigma^2$ is

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} \left(= \frac{\sum_{i=1}^{n} e_i^2}{n-2} \right).$$

Then $E(MSE) = \sigma^2$. MSE is automatically given in SAS and R.

$s = \sqrt{MSE}$ is an estimator of σ, the standard deviation of Y_i. Is it unbiased?

Example: Toluca data. $MSE = 2383.72$ hours^2 and $\sqrt{MSE} = 48.82$ hours from the SAS output.

- So far we have only assumed $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$.
- We can *additionally* assume

$$\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2).$$

- This allows us to make *inference* about β_0, β_1, and obtain prediction intervals for a new Y_h with covariate X_h.
- The model is, succinctly,

$$Y_i \overset{ind.}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2), \quad i = 1, \ldots, n.$$

b_0 and b_1 are MLEs

Fact: Under the assumption of normality, the least squares estimators (b_0, b_1) are also *maximum likelihood estimators* (pp. 27–30) for (β_0, β_1).

The *likelihood* of $(\beta_0, \beta_1, \sigma^2)$ is the density of the data given these parameters (p. 31):

$$
\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma^2) &= f(y_1, \ldots, y_n | \beta_0, \beta_1, \sigma^2) \\
&\overset{ind.}{=} \prod_{i=1}^{n} f(y_i | \beta_0, \beta_1, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-0.5 \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right).
\end{aligned}
$$

LS = MLE under normality

$\mathcal{L}(\beta_0, \beta_1, \sigma^2)$ is maximized when $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ is as small as possible.

$\Rightarrow$ Least-squares estimators are MLEs too!

The MLE of σ^2 is, instead, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$; the denominator changes.

The least squares estimator for the slope is b_1 is

$$b_1 = \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} = \sum_{i=1}^{n} \left[\frac{(X_i - \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right] Y_i .$$

Thus, b_1 is a linear combination n independent normal random variables $Y_1, \ldots, Y_n$. Therefore

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right).$$

We computed $E(b_1) = \beta_1$ before; we use the standard result for the variance of a linear combination of independent random variables for the variance.

$se(b_1)$ estimates $sd(b_1)$

So,

$$\sigma\{b_1\} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Take b_1, subtract off its mean, and divide by its standard deviation and you've got...

$$\frac{b_1 - \beta_1}{\sigma\{b_1\}} \sim N(0, 1).$$

We will never know $\sigma\{b_1\}$; we estimate it by

$$se(b_1) = \sqrt{\frac{MSE}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Question: How do we make $\sigma^2\{b_1\}$ *as small as possible* (p. 50)? If we do this, we cannot actually check the assumption of linearity.

Confidence interval for β_1 and testing $H_0 : \beta_1 = \beta_{10}$

Fact:

$$\frac{b_1 - \beta_1}{se(b_1)} \sim t_{n-2}.$$

A $(1-\alpha)100\%$ CI for β_1 has endpoints

$$b_1 \pm t_{n-2}(1-\alpha/2)se(b_1).$$

Under $H_0 : \beta_1 = \beta_{10}$,

$$t^* = \frac{b_1 - \beta_{10}}{se(b_1)} \sim t_{n-2}.$$

P-values are computed as usual.
Note: Of particular interest is $H_0 : \beta_1 = 0$, that $E(Y_i) = \beta_0$ *and does not depend on* X_i. That is, "H_0: X_i is useless in predicting Y_i."

Table of regression coefficients

Regression output typically produces a table like:

Parameter	Estimate	Standard error	t^*	p-value
Intercept β_0	b_0	$\mathrm{se}(b_0)$	$t_0^* = \frac{b_0}{se(b_0)}$	$P(\|T\| > \|t_0^*\|)$
Slope β_1	b_1	$\mathrm{se}(b_1)$	$t_1^* = \frac{b_1}{se(b_1)}$	$P(\|T\| > \|t_1^*\|)$

where $T \sim t_{n-p}$ and p is the number of parameters used to estimate the mean, here $p = 2$: β_0 and β_1. Later p will be the number of predictors in the model plus one.

The two p-values in the table test $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$ respectively. The test for zero intercept is usually not of interest.

Toluca data

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	62.36586	26.17743	2.38	0.0259
size	Lot Size (parts/lot)	1	3.57020	0.34697	10.29	<.0001

We reject $H_0 : \beta_1 = 0$ at any reasonable significance level ($P < 0.0001$). There is a significant linear association between lot size and hours worked.

Note $se(b_1) = 0.347$, $t_1^* = \frac{3.57}{0.347} = 10.3$, and $P(|t_{23}| > 10.3) < 0.0001$.

We can test non-zero β_1 with a specific form of the `TEST` statement in `PROC REG`. E.g., `slope4:  test size=4;`

2.2 Inference about the intercept β_0

The intercept usually is not very interesting, but just in case...

Write b_0 as a linear combination of $Y_1, \ldots, Y_n$ as we did with the slope:

$$b_0 = \bar{Y} - b_1\bar{X} = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{\sum_{j=1}^{n}(X_j - \bar{X})^2} \right] Y_i.$$

After some slogging, this leads to

$$b_0 \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right]\right).$$

Distribution of $\frac{b_0-\beta_0}{se(b_0)}$

Define $\mathrm{se}(b_0) = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i-\bar{X})^2}\right]}$ and you're in business:

$$\frac{b_0 - \beta_0}{se(b_0)} \sim t_{n-2}.$$

Obtain CIs and tests about β_0 as usual...

Estimating $E(Y_h) = \beta_0 + \beta_1 X_h$
(e.g. inference about the regression line)
Let X_h be *any predictor*; say we want to estimate the mean of all outcomes in the *population* that have covariate X_h. This is given by

$$E(Y_h) = \beta_0 + \beta_1 X_h.$$

Our estimator of this is

$$\begin{aligned}
\hat{Y}_h &= b_0 + b_1 X_h \\
&= \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{\sum_{j=1}^{n}(X_j - \bar{X})^2} + \frac{(X_i - \bar{X})X_h}{\sum_{j=1}^{n}(X_j - \bar{X})^2} \right] Y_i \\
&= \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(X_h - \bar{X})(X_i - \bar{X})}{\sum_{j=1}^{n}(X_j - \bar{X})^2} \right] Y_i
\end{aligned}$$

Distribution of $\hat{Y}_h$

Again we have a linear combination of independent normals as our estimator. This leads, after slogging through some math (pp. 53–54), to

$$b_0 + b_1 X_h \sim N\left(\beta_0 + \beta_1 X_h, \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]\right).$$

As before, this leads to a $(1-\alpha)100\%$ CI for $\beta_0 + \beta_1 X_h$

$$b_0 + b_1 X_h \pm t_{n-2}(1-\alpha/2) se(b_0 + b_1 X_h),$$

where $\text{se}(b_0 + b_1 X_h) = \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]}$.
Question: For what value of x_h is the CI narrowist? What happens when X_h moves away from $\bar{X}$?

2.5 Prediction intervals

- We discussed constructing a CI for the unknown mean at X_h, $\beta_0 + \beta_1 X_h$.
- What if we want to find an interval that contains a single Y_h with fixed probability?
- If we knew β_0, β_1, and σ^2 this is easy:

$$Y_h = \beta_0 + \beta_1 X_h + \epsilon_h,$$

and so, for example,

$$P(\beta_0 + \beta_1 X_h - 1.96\sigma \leq Y_h \leq \beta_0 + \beta_1 X_h + 1.96\sigma) = 0.95.$$

- Unfortunately, we don't know β_0 and β_1. We don't even know σ, but we can construct a random variable with a t distribution to develop an appropriate *prediction interval*.

Variability of $Y_h - \hat{Y}_h$

An interval that contains Y_h (independent of $Y_1, \ldots, Y_n$) with $(1-\alpha)$ probability needs to account for

1. The variability of the least squares line $b_0 + b_1 X_h$, and
2. The natural variability of response Y_h built into the model; $\epsilon_h \sim N(0, \sigma^2)$.

We have

$$\begin{aligned} \sigma^2\left\{Y_h - \hat{Y}_h\right\} &\overset{ind}{=} \sigma^2\{Y_h\} + \sigma^2\left\{\hat{Y}_h\right\} \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right] \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right] \end{aligned}$$

Prediction interval

Since $Y_h - \hat{Y}_h \sim N\left(0, \sigma^2\left\{Y_h - \hat{Y}_h\right\}\right)$,

$$\frac{Y_h - \hat{Y}_h}{\hat{\sigma}\left\{Y_h - \hat{Y}_h\right\}} \sim t_{n-2}$$

We thus obtain a $(1-\alpha/2)100\%$ *prediction interval* (PI) for Y_h:

$$b_0 + b_1 X_h \pm t_{n-2}(1-\alpha/2)\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]}.$$

Note: As $n \to \infty$, $b_0 \xrightarrow{P} \beta_0$, $b_1 \xrightarrow{P} \beta_1$, $t_{n-2}(1-\alpha/2) \to \Phi^{-1}(1-\alpha/2)$, and $MSE \xrightarrow{P} \sigma^2$. That is, as the sample size grows, the prediction interval converges to

$$\beta_0 + \beta_1 x_h \pm \Phi^{-1}(1-\alpha/2)\sigma.$$

Example: Toluca data

- Find a 95% CI for the mean number of work hours for lots of size $X_h = 65$ units.
- Find a 95% PI for the number of work hours for a lot of size $X_h = 65$ units.
- Repeat both for $X_h = 100$ units.
- SAS code follows...

SAS code

```
data toluca;
input size hours @@;
label size="Lot Size (parts/lot)";
label hours="Work Hours";
datalines;
   80  399   30  121   50  221   90  376   70  361   60  224  120  546
   80  352  100  353   50  157   40  160   70  252   90  389   20  113
  110  435  100  420   30  212   50  268   90  377  110  421   30  273
   90  468   40  244   80  342   70  323
;
proc sql;
insert into toluca
(size)
values(65)
values(100);
quit;

options nocenter;
proc reg data=toluca;
 model hours=size / clm cli alpha=0.05;
run;
```

SAS output

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	399.0000	347.9820	10.3628	326.5449	369.4191	244.7333	451.2307	51.0180
2	121.0000	169.4719	16.9697	134.3673	204.5765	62.5464	276.3975	-48.4719
3	221.0000	240.8760	11.9793	216.0948	265.6571	136.8815	344.8704	-19.8760
				...et cetera...				
24	342.0000	347.9820	10.3628	326.5449	369.4191	244.7333	451.2307	-5.9820
25	323.0000	312.2800	9.7647	292.0803	332.4797	209.2811	415.2789	10.7200
26	.	294.4290	9.9176	273.9129	314.9451	191.3676	397.4904	.
27	.	419.3861	14.2723	389.8615	448.9106	314.1604	524.6117	.

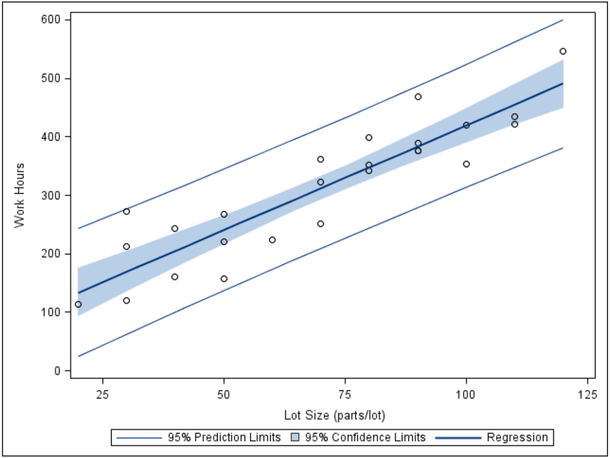
More SAS code & output

```
proc reg data=toluca;
 model hours=size / clm cli alpha=0.05;
 output out=regstats lclm=lclm uclm=uclm lcl=lcl ucl=ucl p=pred r=r;
run;

proc print data=regstats;
 var hours size lclm uclm lcl ucl pred;
run;
```

```
Obs    hours    size     lclm      uclm      lcl       ucl       pred
  1     399      80    326.545   369.419   244.733   451.231   347.982
  2     121      30    134.367   204.577    62.546   276.397   169.472
  3     221      50    216.095   265.657   136.882   344.870   240.876
                              ...et cetera...
 24     342      80    326.545   369.419   244.733   451.231   347.982
 25     323      70    292.080   332.480   209.281   415.279   312.280
 26       .      65    273.913   314.945   191.368   397.490   294.429
 27       .     100    389.862   448.911   314.160   524.612   419.386
```

SAS plot of 95% CI for mean & prediction intervals

```
proc sgplot data=toluca;
 reg x=size y=hours / clm cli;
run;
```



Obtaining confidence intervals for β_0 and β_1

SAS code:

```
options nocenter;
proc reg data=toluca;
 model hours=size / clb alpha=0.01;
run;
```

Output:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	Intercept	1	62.36586	26.17743	2.38	0.0259	-11.12299	135.85470
size	Lot Size (parts/lot)	1	3.57020	0.34697	10.29	<.0001	2.59613	4.54427

2.6 Credible band for regression function

- Gives *region that entire regression line lies in* with certain probability/confidence.
- Given by

$$\hat{Y}_h \pm W \; se\{\hat{Y}_h\} = b_0 + b_1 X_h \pm W \; se\{b_0 + b_1 X_h\}$$

where $W^2 = 2F(1-\alpha; 2, n-2)$

- Defined for $X_h \in \mathbb{R}$. Ignore for nonsense values of X_h.
- Not straightforward to get in SAS (or other packages).