Chapters 1 and 2

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I
Toluca makes replacement parts for refrigerators.

We consider one particular part, manufactured in varying lot sizes.

Takes time to set up production regardless of lot size; this time plus machining & assembly makes up work hours.

Want to relate work hours to lot size.

\[ n = 25 \text{ pairs } (x_i, Y_i) \text{ were obtained.} \]
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 sgscatter; plot hours*size; run;
options nocenter;
proc reg; model hours=size; run;
The REG Procedure
Dependent Variable: hours Work Hours

Number of Observations Read 25
Number of Observations Used 25

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>252378</td>
<td>252378</td>
<td>105.88</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>23</td>
<td>54825</td>
<td>2383.71562</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>24</td>
<td>307203</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Parameter Estimates

| 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 |
Roughly linear trend, no obvious outliers.
The fitted model is

\[
\hat{\text{hours}} = 62.37 + 3.570 \text{ lot size}.
\]

- A lot size of \( x = 65 \) takes \( \hat{Y} = 62.37 + 3.570(65) = 294 \) 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?
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 $\epsilon_i$. 

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 \((\bar{x}, \bar{Y})\) (easy to show).
\(\sigma^2\) is the error variance. If we observed the \(\epsilon_1, \ldots, \epsilon_n\), a natural estimator is \(S^2 = \frac{1}{n} \sum_{i=1}^{n} (\epsilon_i - 0)^2\). If we replace each \(\epsilon_i\) by \(e_i\) we have \(\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2\). However,

\[
E(\hat{\sigma}^2) = \frac{1}{n} \sum_{i=1}^{n} E(Y_i - b_0 - b_1 x_i)^2
\]

\[= \ldots \text{a lot of hideous algebra later...}\]

\[= \frac{n - 2}{n} \sigma^2.\]

So in the end we use the unbiased \textit{mean squared error}

\[
MSE = \frac{1}{n - 2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n - 2} \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2.
\]
So an estimate of $\text{var}(Y_i) = \sigma^2$ is

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

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

$s = \sqrt{MSE}$ is an estimator of $\sigma$, 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 $\beta_0$, $\beta_1$, and obtain prediction intervals for a new $Y_h$ with covariate $x_h$.

The model is, succinctly,

$$Y_i \overset{ind.}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, \ldots, n.$$
Fact: Under the assumption of normality, the least squares estimators \((b_0, b_1)\) are also maximum likelihood estimators (pp. 27–30) for \((\beta_0, \beta_1)\).

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

\[
\mathcal{L}(\beta_0, \beta_1, \sigma^2) = f(y_1, \ldots, y_n|\beta_0, \beta_1, \sigma^2)
\]

\[
\overset{\text{ind.}}{=} \prod_{i=1}^{n} f(y_i|\beta_0, \beta_1, \sigma^2)
\]

\[
= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -0.5 \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \right)
\]

\[
= (2\pi\sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right).
\]
$\mathcal{L}(\beta_0, \beta_1, \sigma^2)$ 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 $\sigma^2$ 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,

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

Take \( b_1 \), subtract off its mean, and divide by its standard deviation and you’ve got...

\[
\frac{b_1 - \beta_1}{sd(b_1)} \sim N(0, 1).
\]

We will never know \( sd(b_1) \); we estimate it by

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

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

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

A \((1 - \alpha)\)100\% CI for \(\beta_1\) 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 \(Y_i\)”
Regression output typically produces a table like:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard error</th>
<th>$t^*$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept $\beta_0$</td>
<td>$b_0$</td>
<td>se($b_0$)</td>
<td>$t_0^* = \frac{b_0}{se(b_0)}$</td>
<td>$P(</td>
</tr>
<tr>
<td>Slope $\beta_1$</td>
<td>$b_1$</td>
<td>se($b_1$)</td>
<td>$t_1^* = \frac{b_1}{se(b_1)}$</td>
<td>$P(</td>
</tr>
</tbody>
</table>

where $T \sim t_{n-p}$ and $p$ is the number of parameters used to estimate the mean, here $p = 2$: $\beta_0$ and $\beta_1$. 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 | 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 |

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

$$b_0 \sim \mathcal{N} \left( \beta_0, \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \right).$$
Define $se(b_0) = \sqrt{MSE \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]}$ and you’re in business:

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

Obtain CIs and tests about $\beta_0$ as usual...
2.4 Estimating $E(Y_h)$

**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

$$\hat{Y}_h = b_0 + b_1 x_h$$

$$= \sum_{i=1}^{n} \left[ \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} + \frac{(x_i - \bar{x})x_h}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right] Y_i$$

$$= \sum_{i=1}^{n} \left[ \frac{1}{n} + \frac{(x_h - \bar{x})(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right] Y_i$$
Again we have a linear combination of independent normals as our estimator. This leads, after slogging through some math (pp. 53–54), to

\[
b_0 + b_1x_h \sim N \left( \beta_0 + \beta_1x_h, \sigma^2 \left[ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \right).
\]

As before, this leads to a \((1 - \alpha)100\%\) CI for \(\beta_0 + \beta_1x_h\)

\[
b_0 + b_1x_h \pm t_{n-2}(1 - \alpha/2)se(b_0 + b_1x_h),
\]

where \(se(b_0 + b_1x_h) = \sqrt{MSE \left[ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]}\).

**Question:** For what value of \(x_h\) is the CI narrowest? What happens when \(x_h\) moves away from \(\bar{x}\)?
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 the actual value $Y_h$ is in (versus only it’s mean) with fixed probability?

If we knew $\beta_0$, $\beta_1$, and $\sigma^2$ 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 $\beta_0$ and $\beta_1$. We don’t even know $\sigma$, but we can estimate all three of these.
Variability of $b_0 + b_1 x_h + \epsilon_h$

An interval that contains $Y_h$ with $(1 - \alpha)$ probability needs to account for

1. The variability of the estimators $b_0$ and $b_1$; i.e. we don’t know exactly where $\beta_0 + \beta_1 x_h$ is, and

2. The natural variability of response $Y_h$ built into the model; $\epsilon_h \sim N(0, \sigma^2)$.

We have

$$\text{var}(b_0 + b_1 x_h + \epsilon_h) = \text{var}(b_0 + b_1 x_h) + \text{var}(\epsilon_h)$$

$$= \sigma^2 \left[ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] + \sigma^2$$

$$= \sigma^2 \left[ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 + 1} \right]$$
Estimating $\sigma^2$ by MSE we obtain a $(1 - \alpha/2)100\%$ prediction interval (PI) for $Y_h$ is

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

**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...
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
    ;
data predict;
  input size hours;
  datalines;
    65 .
    100 .
    ;

data toluca;
  set toluca predict;

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

<table>
<thead>
<tr>
<th>Obs</th>
<th>Variable</th>
<th>Value</th>
<th>Mean</th>
<th>Predict</th>
<th>Std Error</th>
<th>95% CL Mean</th>
<th>95% CL Predict</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>399.0000</td>
<td>347.9820</td>
<td>10.3628</td>
<td>10.3628</td>
<td>326.5449</td>
<td>369.4191</td>
<td>244.7333</td>
<td>451.2307</td>
</tr>
<tr>
<td>2</td>
<td>121.0000</td>
<td>169.4719</td>
<td>16.9697</td>
<td>16.9697</td>
<td>134.3673</td>
<td>204.5765</td>
<td>62.5464</td>
<td>276.3975</td>
</tr>
<tr>
<td>3</td>
<td>221.0000</td>
<td>240.8760</td>
<td>11.9793</td>
<td>11.9793</td>
<td>216.0948</td>
<td>265.6571</td>
<td>136.8815</td>
<td>344.8704</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>24</td>
<td>342.0000</td>
<td>347.9820</td>
<td>10.3628</td>
<td>10.3628</td>
<td>326.5449</td>
<td>369.4191</td>
<td>244.7333</td>
<td>451.2307</td>
</tr>
</tbody>
</table>
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;
```

<table>
<thead>
<tr>
<th>Obs</th>
<th>hours</th>
<th>size</th>
<th>lclm</th>
<th>uclm</th>
<th>lcl</th>
<th>ucl</th>
<th>pred</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>399</td>
<td>80</td>
<td>326.545</td>
<td>369.419</td>
<td>244.733</td>
<td>451.231</td>
<td>347.982</td>
</tr>
<tr>
<td>2</td>
<td>121</td>
<td>30</td>
<td>134.367</td>
<td>204.577</td>
<td>62.546</td>
<td>276.397</td>
<td>169.472</td>
</tr>
<tr>
<td>3</td>
<td>221</td>
<td>50</td>
<td>216.095</td>
<td>265.657</td>
<td>136.882</td>
<td>344.870</td>
<td>240.876</td>
</tr>
<tr>
<td>24</td>
<td>342</td>
<td>80</td>
<td>326.545</td>
<td>369.419</td>
<td>244.733</td>
<td>451.231</td>
<td>347.982</td>
</tr>
<tr>
<td>25</td>
<td>323</td>
<td>70</td>
<td>292.080</td>
<td>332.480</td>
<td>209.281</td>
<td>415.279</td>
<td>312.280</td>
</tr>
<tr>
<td>26</td>
<td>65</td>
<td>273.913</td>
<td>314.945</td>
<td>191.368</td>
<td>397.490</td>
<td>294.429</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>100</td>
<td>389.862</td>
<td>448.911</td>
<td>314.160</td>
<td>524.612</td>
<td>419.386</td>
<td></td>
</tr>
</tbody>
</table>
proc sgplot data=toluca;
   reg x=size y=hours / clm cli;
run;
Obtaining confidence intervals for $\beta_0$ and $\beta_1$

SAS code:

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

Output:

| Variable  | Label                  | DF | Parameter   | Standard   | 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   | < 0.0001| 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

\[
\hat{Y}_h \pm W \text{ se}\{\hat{Y}_h\} = b_0 + b_1x_h \pm W \text{ se}\{b_0 + b_1x_h\}
\]

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).