Simple Linear Regression (Chapter 1 & 2)

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

When to Use What Statistics

Statistical Analyses	Independent Variables		Dependent Variables		Control Variables
	# of IVs	Data Type	# of DVs	Type of Data	
Chi square	1	categorical	1	categorical	0
***t*-Test**	1	dichotomous	1	continuous	0
ANOVA	1 +	categorical	1	continuous	0
ANCOVA	1 +	categorical	1	continuous	1 +
MANOVA	1 +	categorical	2 +	continuous	0
MANCOVA	1 +	categorical	2 +	continuous	1 +
Correlation	1	dichotomous or continuous	1	continuous	0
Multiple regression	2 +	dichotomous or continuous	1	continuous	0
Path analysis	2 +	continuous	1 +	continuous	0
Logistic Regression	1 +	categorical or continuous	1	dichotomous	0

DV: dependent variable, response variable, outcome, phenotype (Y)
IV: independent variable, predictor variable, covariate (X)
Does the difference in gene expression exist between patients with/without a mutation?

When to Use What Statistics

Statistical Analyses	Independent Variables		Dependent Variables		Control Variables
	# of IVs	Data Type	# of DVs	Type of Data	
Chi square	1	categorical	1	categorical	0
***t*-Test**	1	dichotomous	1	continuous	0
ANOVA	1 +	categorical	1	continuous	0
ANCOVA	1 +	categorical	1	continuous	1 +
MANOVA	1 +	categorical	2 +	continuous	0
MANCOVA	1 +	categorical	2 +	continuous	1 +
Correlation	1	dichotomous or continuous	1	continuous	0
Multiple regression	2 +	dichotomous or continuous	1	continuous	0
Path analysis	2 +	continuous	1 +	continuous	0
Logistic Regression	1 +	categorical or continuous	1	dichotomous	0

DV: dependent variable, response variable, outcome, phenotype (Y)
IV: independent variable, predictor variable, covariate (X)
Does the difference in gene expression exist between patients with/without a mutation?
Determine the association between disease status (Yes, No) and genotype (AA, Aa, aa).

When to Use What Statistics

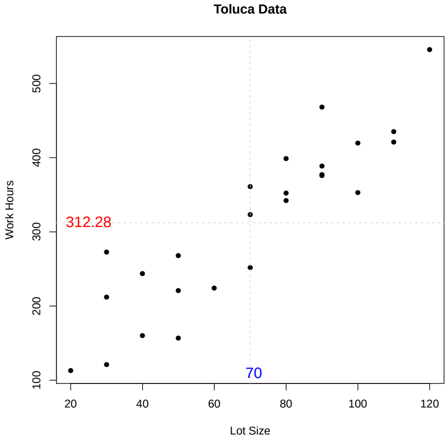
Statistical Analyses	Independent Variables		Dependent Variables		Control Variables
	# of IVs	Data Type	# of DVs	Type of Data	
Chi square	1	categorical	1	categorical	0
***t*-Test**	1	dichotomous	1	continuous	0
ANOVA	1 +	categorical	1	continuous	0
ANCOVA	1 +	categorical	1	continuous	1 +
MANOVA	1 +	categorical	2 +	continuous	0
MANCOVA	1 +	categorical	2 +	continuous	1 +
Correlation	1	dichotomous or continuous	1	continuous	0
Multiple regression	2 +	dichotomous or continuous	1	continuous	0
Path analysis	2 +	continuous	1 +	continuous	0
Logistic Regression	1 +	categorical or continuous	1	dichotomous	0

DV: dependent variable, response variable, outcome, phenotype (Y)
IV: independent variable, predictor variable, covariate (X)
Does the difference in gene expression exist between patients with/without a mutation?
Determine the association between disease status (Yes, No) and genotype (AA, Aa, aa).
Predict daughter's height from father's height.

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



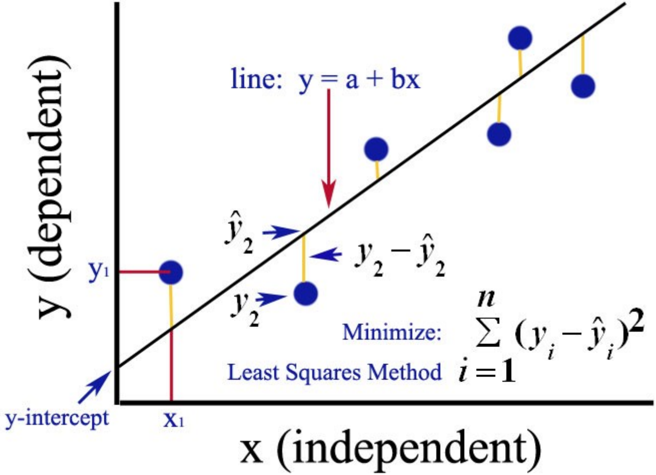
Roughly linear trend, no obvious outliers.

The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i the value of the response variable in the i^{th} trial
- β_0, β_1 are parameters
- X_i is known; it is the value of the predictor variable in the i^{th} trial
- ϵ_i is a random error term with $E(\epsilon_i) = 0$ and finite variance $\sigma^2(\epsilon_i) = \sigma^2$
- $i = 1, 2, ...n$
- $\hat{Y} = E(Y_i) = \beta_0 + \beta_1 X_i$

Least Square Linear Regression



Least Squares Estimation

Seek to minimize

$$Q = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

Minimize by maximizing -Q.

$$\frac{dQ}{d\beta_0} = 0$$

$$\frac{dQ}{d\beta_1} = 0$$

Normal Equations

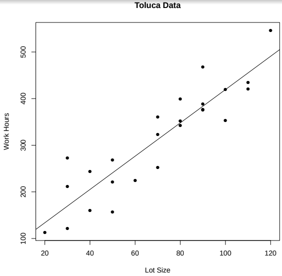
The result of this maximization step are called the normal equations.

$$
\begin{aligned}
\sum Y_i &= nb_0 + b_1 \sum X_i \\
\sum X_i Y_i &= b_0 \sum X_i + b_1 \sum X_i^2
\end{aligned}
$$

The solution to the normal equations:

$$
\begin{aligned}
b_1 &= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \\
b_0 &= \overline{Y} - b_1 \overline{X}
\end{aligned}
$$

Toluca

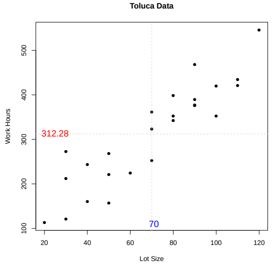


The fitted model is

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

- 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? (We don't observe any data with lot size $=0$)

Alternative Model: Centering



$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\
Y_i &= \beta_0^* + \beta_1 (X_i - \overline{X}) + \epsilon_i
\end{aligned}
$$

$$
\begin{aligned}
\hat{Y} &= 62.37 + 3.570X \\
\hat{Y} &= 312.28 + 3.570(X - 70)
\end{aligned}
$$

- $b_0^* = b_0 + b_1\overline{X} = \overline{Y}$.
- β_0^* is the mean outcome when $X = 70$ (reference group).
- Interpretation for β_1 has not changed.

R Code

```
>fit1<-lm(dat[,2] ~ dat[,1])
>summary(fit1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
dat[, 1]       3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215,Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10

>xstar<-dat[,1]-mean(dat[,1])
>fit2<-lm(dat[,2] ~ xstar)
>summary(fit2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  312.280      9.765   31.98  < 2e-16 ***
xstar          3.570      0.347   10.29 4.45e-10 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215,Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

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.

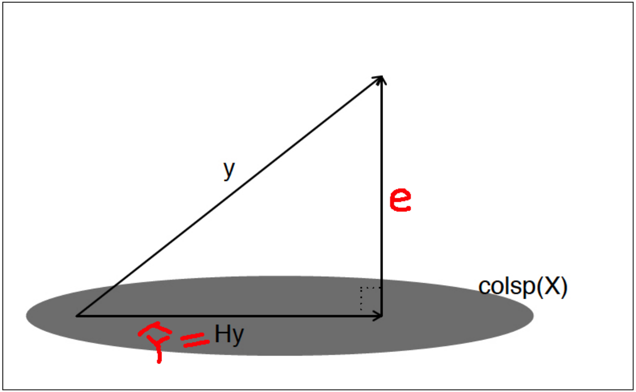
Properties of the residuals (pp. 23–24)

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 $(\bar{X}, \bar{Y})$.

Plug in $\overline{X}$ in the model

$$\begin{aligned}\hat{Y}_i &= b_0 + b_1 X_i \\ \hat{Y}_i &= \overline{Y} - b_1 \overline{X} + b_1 \overline{X}\end{aligned}$$

Orthogonal projection of 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.

Example: Toluca data. $MSE = 2383.72$ hours^2 and $\sqrt{MSE} = 48.82$ hours from the R output. For a lot size of $X = 65$ units, the mean work hour ($\hat{Y}$) is 294.4 hours. The variation in work hours from lot to lot for lots of 65 units is quite substantial since the prediction would still be off by $\frac{48.82}{294.4} \approx 16.6\%$.

- 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):

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

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.

Section 2.1: Inferences on β_1

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

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

(proof in pp. 43)

$se(b_1)$ 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

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

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	$\text{se}(b_0)$	$t_0^* = \frac{b_0}{se(b_0)}$	$P(\lvert T\rvert > \lvert t_0^*\rvert)$
Slope β_1	b_1	$\text{se}(b_1)$	$t_1^* = \frac{b_1}{se(b_1)}$	$P(\lvert T\rvert > \lvert t_1^*\rvert)$

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 $P(|t_{23}| > 10.3) < 0.0001$.

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 $Y_1, \dots, Y_n$ 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 $\text{se}(b_0) = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}$ 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...

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

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

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

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

This is different from the CI for $\hat{Y}_h$ (mean). The prediction interval of next data point (Y_h, not the mean) includes the uncertainty in the population mean, plus data scatter. So a prediction interval is always wider than a confidence interval.

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:

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

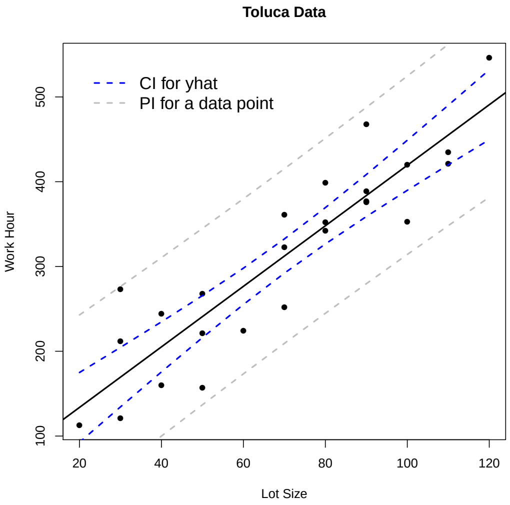
Note: As $n \to \infty$, $b_0 \xrightarrow{P} \beta_0$, $b_1 \xrightarrow{P} \beta_1$, $t_{n-2}(1-\alpha/2) \to \Phi^{-1}(1-\alpha/2)$, and $MSE \xrightarrow{P} \sigma^2$. That is, as the sample size grows, the prediction interval converges to

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

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.
- R Code in Lab6.R

Plot of 95% CI for mean & prediction intervals



Obtaining confidence intervals for β_0 and β_1

R code:

```
> confint(fit1)
                2.5 %     97.5 %
(Intercept) 8.213711 116.518006
LotSize     2.852435   4.287969
```

2.6 Confidence 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.
- R code in Lab6.R

Confidence band for regression function