Simple Linear Regression: A Model for the Mean

HYPOTHESIS TESTS

CONFIDENCE INTERVALS

(IGNORE: “CALIBRATION INTERVALS” AND “PLANNING AN EXPERIMENT: REPLICATION”)

Notation for the Mean

- Y is the response variable
- X is the explanatory variable
- $\mu\{Y|X\}$ is the "mean of Y as a function of X"

For Simple Linear Regression (SLR), we write this mean as

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

- β_0 has the same **units** as Y

(this is the intercept)

- β_1 has the same **units** as Y/ X

(this is a rate or slope)

Example: Y (deaths per million) is mortality from skin cancer in a state & X is state latitude (in degrees)

β_0 is in deaths per million
β_1 is in (deaths per million)/degrees

Confidence Intervals and Hypothesis Tests

The current chapter focusses on 4 major confidence intervals/tests:

- For β_0
- For β_1

We will cover these first

- For the mean value of Y at X_0, $\mu\{Y|X_0\} = \beta_0 + \beta_1 X_0$
- For a prediction of Y at X_0, $\text{Pred}\{Y|X_0\}$

The Estimators

Movie Budgets and Gross Find the best predicted gross amount for a movie with a budget of 40 million dollars. (In the table below, all amounts are in millions of dollars.)

Budget	62	90	50	35	200	100	90
Gross	65	64	48	57	601	146	47

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}}$$

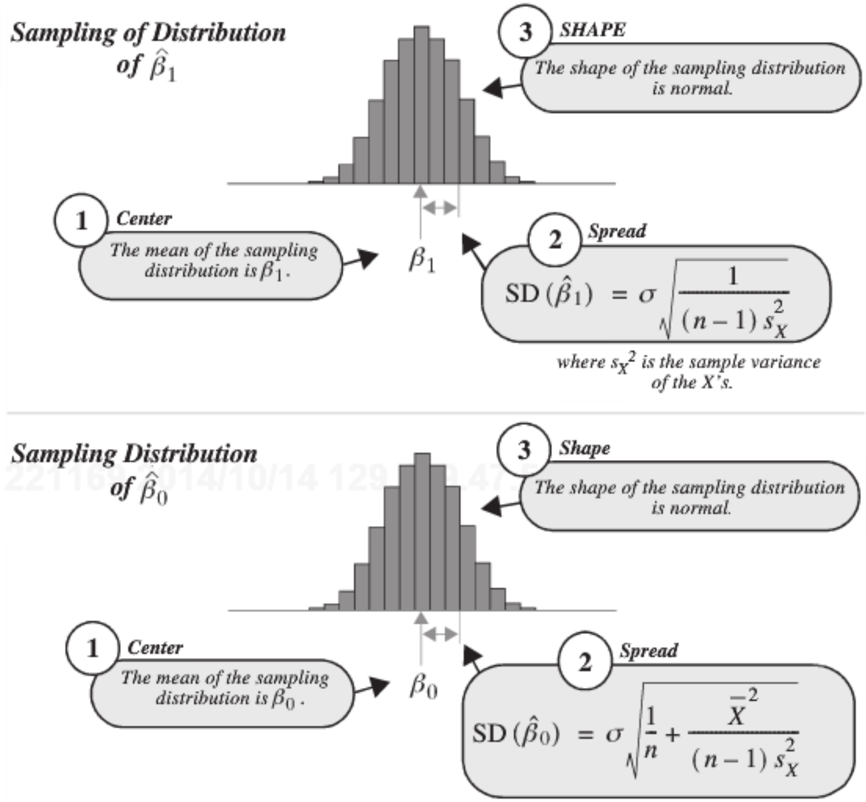
$\bar{X} = 89.57$ $\bar{Y} = 146.86$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-164.14293	65.06146	-2.52	0.0530
budget	1	3.47209	0.63378	5.48	0.0028

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

Sampling Distributions & Hypothesis Test



$$\hat{\sigma} = \sqrt{\frac{\textit{Sum of all squared residuals}}{\textit{Degrees of freedom}}},$$

$$\mathrm{SE}(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1){s_X}^2}}, \quad \text{d.f.} = n-2$$

$$\mathrm{SE}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1){s_X}^2}}, \quad \text{d.f.} = n-2,$$

where ${s_X}^2$ is the sample variance of the X's.

Two Hypothesis Tests:

$H_0\colon \beta_0 = 0$	$H_0\colon \beta_1 = 0$
$H_A\colon \beta_0 \neq 0$	$H_A\colon \beta_1 \neq 0$

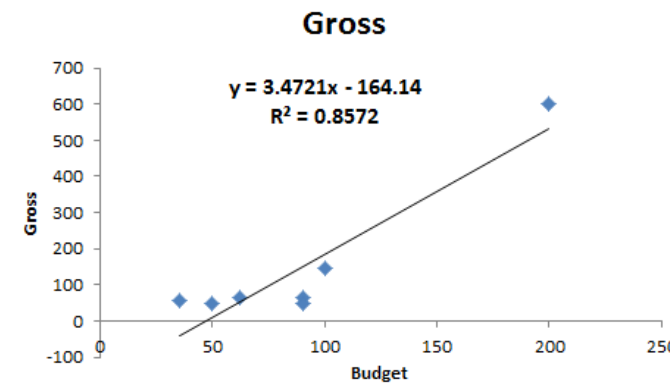
The Estimators

Movie Budgets and Gross Find the best predicted gross amount for a movie with a budget of 40 million dollars. (In the table below, all amounts are in millions of dollars.)

Budget	62	90	50	35	200	100	90
Gross	65	64	48	57	601	146	47

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-164.14293	65.06146	-2.52	0.0530
budget	1	3.47209	0.63378	5.48	0.0028



Estimate $\pm$ $t_{\alpha/2,df}$*SE (df = n-2)

Intercept $\pm$ $t_{.025,5}$*SE

-164.143 $\pm$ 2.571*65.06

(-331.39,3.103)

We estimate that budget of $0 is associated with a gross between $0 and $3.103M (95% CI)

Estimate $\pm$ $t_{\alpha/2,df}$*SE

Budget_slope $\pm$ $t_{.025,5}$*SE

3.472 $\pm$ 2.571*.6338

(1.84, 5.10)

We estimate that an increase in budget of $1 million is associated with an increase in gross between $1.84M and $5.10M

Confidence Intervals and Hypothesis Tests

The current chapter focusses on 4 major confidence intervals/tests:

Now these

- For β_0
- For β_1

(Confidence Intervals)

- For the mean value of Y at X_0, $\mu\{Y|X_0\} = \beta_0 + \beta_1 X_0$
- For a prediction of Y at X_0, $\mathrm{Pred}\{Y|X_0\}$

(Prediction Intervals)

- Both will be based around our estimate of μ: $\hat{\mu}\{Y|X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$

Confidence Intervals & Prediction Intervals

$$\mathrm{SE}[\hat{\mu}\{Y|X_0\}] = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_X^{\,2}}},$$

(Confidence Interval)

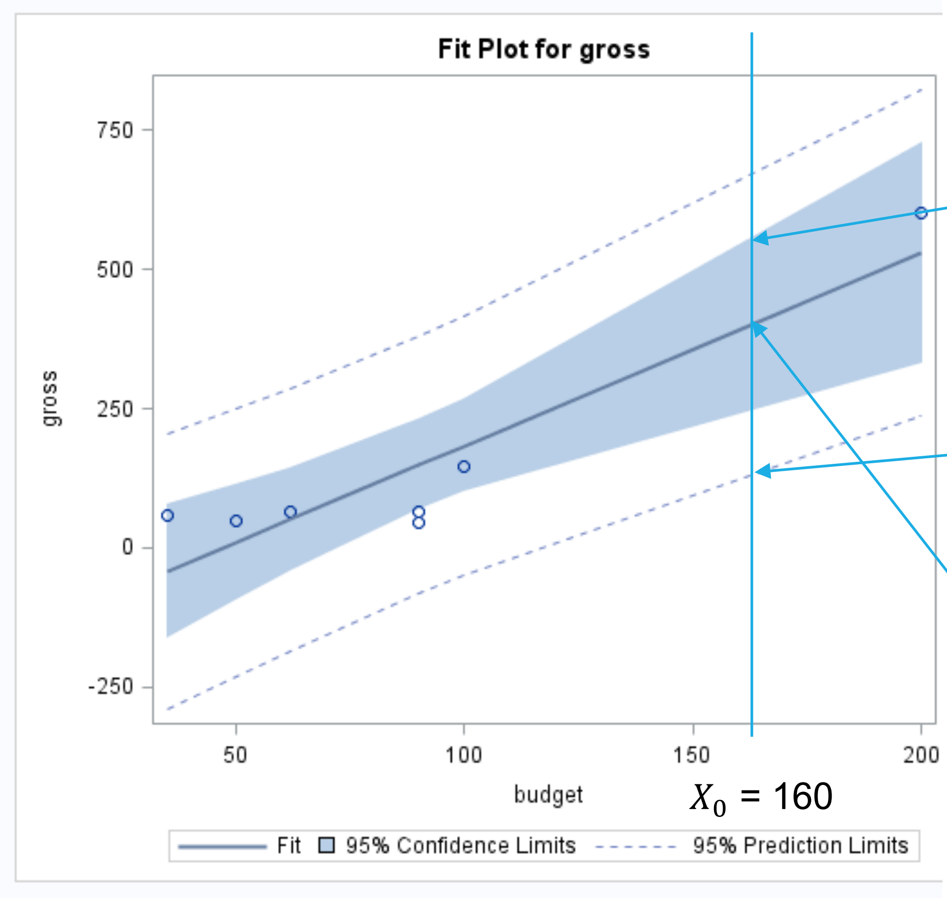
$$\mathrm{SE}[\mathrm{Pred}\{Y|X_0\}] = \sqrt{\hat{\sigma}^2 + \mathrm{SE}[\hat{\mu}\{Y|X_0\}]^2}$$

Prediction error = *Random sampling error* + *Estimation error*

$$\mathrm{SE}[\mathrm{Pred}\{Y|X_0\}] = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_X^{\,2}}}$$

(Prediction intervals will always be wider and won't go to zero as $n \to \infty$)

Confidence Intervals & Prediction Intervals



X_0 = 160

$$\mathrm{SE}[\hat{\mu}\{Y|X_0\}] = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_X{}^2}},$$

$$\mathrm{SE}[\mathrm{Pred}\{Y|X_0\}] = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_X{}^2}}$$

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

CI: $\hat{\mu}\{Y|X_0\} \pm t_{.025,n-2}$*SE

A Few Additional Topics

Differing Terminology & Causation

You will hear or read about alternative terms for “explanatory” and “response” variables:

Explanatory:
~~Independent variable~~
Exogenous variable
Predictor variable
Covariate
Feature
Input

Response:
~~Dependent variable~~
Endogenous variable
Supervisor
Output

It is dangerously tempting to interpret regression as X **causing** Y even with an observational study → use “association”

Correlation

The **sample correlation coefficient** describes the “degree of linear association between X and Y”

It is commonly denoted “r” and must be between -1 and 1

It is symmetric with respect to X and Y (unlike regression)

Often, we write $R^2 = r^2$ instead which is between 0 and 1

(Interpretation: R^2 is the proportion of the total variation in Y explained by it’s least squares fit on X)

Correlation

Note the correlation will be part of a broader model checking procedure

It is important to not read too much into it as a single indicator of model fit