2. Inference in Regression Analysis

Linear combinations of normals are normal

Fact 4: If $Y_i \sim N(\mu_i, \sigma_i^2)$ and Y_i's are independent, and $a_1, a_2, \ldots, a_n$ are known constants then

$$\sum_{i=1}^{n} a_i Y_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

Thus, a linear combination of independent normal random variables is itself a normal random variable.

Theorem:
b_0 and b_1 are linear combinations of the Y_i's. That is, we can write

$$b_1 = \sum_{i=1}^{n} k_i Y_i \quad \text{and} \quad b_0 = \sum_{i=1}^{n} l_i Y_i$$

where $k_1, \ldots, k_n$ and $l_1, \ldots, l_n$ are known constants.

Proof: Recall $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$. So

$$\begin{aligned}
b_1 &= \frac{1}{S_{XX}} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \frac{1}{S_{XX}} \left[\sum_{i=1}^{n} (X_i - \bar{X}) Y_i - \bar{Y} \sum_{i=1}^{n} (X_i - \bar{X}) \right] \\
&= \frac{1}{S_{XX}} \sum_{i=1}^{n} (X_i - \bar{X}) Y_i = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_{XX}} \right) Y_i \\
&= \sum_{i=1}^{n} k_i Y_i \quad \text{with} \quad k_i = \frac{X_i - \bar{X}}{S_{XX}}
\end{aligned}$$

$$\begin{aligned}
b_0 &= \bar{Y} - b_1 \bar{X} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \bar{X} \sum_{i=1}^{n} k_i Y_i \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - k_i \bar{X} \right) Y_i \\
&= \sum_{i=1}^{n} l_i Y_i \quad \text{with} \quad l_i = \frac{1}{n} - k_i \bar{X}.
\end{aligned}$$

Thus, b_0 and b_1 are linear combinations of the Y_i's and, hence (with fact 4), they are normal variates. What about their means and variances?

Theorem: Under SLR model with normal errors:

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right) \quad \text{and} \quad b_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\frac{\sum_i X_i^2}{S_{XX}}\right).$$

We are first interested in $\sum_i k_i$, $\sum_i k_i X_i$ and $\sum_i k_i^2$.

$$\begin{aligned}
\sum_{i=1}^n k_i &= \sum_{i=1}^n \frac{X_i - \bar{X}}{S_{XX}} = \frac{1}{S_{XX}}\sum_{i=1}^n (X_i - \bar{X}) = 0 \\
\sum_{i=1}^n k_i X_i &= \sum_{i=1}^n \frac{X_i - \bar{X}}{S_{XX}} X_i = \frac{1}{S_{XX}} S_{XX} = 1 \\
\sum_{i=1}^n k_i^2 &= \frac{1}{S_{XX}^2}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{S_{XX}}.
\end{aligned}$$

Proof: Since $b_1 = \sum_{i=1}^n k_i Y_i$, we get

$$E(b_1) = \sum_{i=1}^n k_i E(Y_i) = \sum_{i=1}^n k_i(\beta_0 + \beta_1 X_i).$$

Because $\sum_i k_i = 0$ and $\sum_i k_i X_i = 1$, this is

$$E(b_1) = \beta_0 \sum_{i=1}^n k_i + \beta_1 \sum_{i=1}^n k_i X_i = \beta_1.$$

With $\sum_i k_i^2 = 1/S_{XX}$, we get

$$\begin{aligned}
Var(b_1) &= Var\left(\sum_{i=1}^n k_i Y_i\right) = \sum_{i=1}^n k_i^2 Var(Y_i) \\
&= \sigma^2 \sum_{i=1}^n k_i^2 = \frac{\sigma^2}{S_{XX}}.
\end{aligned}$$

Showing $b_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\frac{\sum_i X_i^2}{S_{XX}}\right)$ is basically the same.

Example: 93 house prices in G'ville sold Dec. 1995.
$Y =$ selling price (in 1,000\$), $X =$ area (1,000 sq.foot)

Assume the SLR model holds

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

LS estimators are $b_0 = -25.2$ and $b_1 = 75.6$. We are interested in testing

$H_0 : \beta_1 = 0$ (no linear relation between area and price)
$H_A : \beta_1 \neq 0$

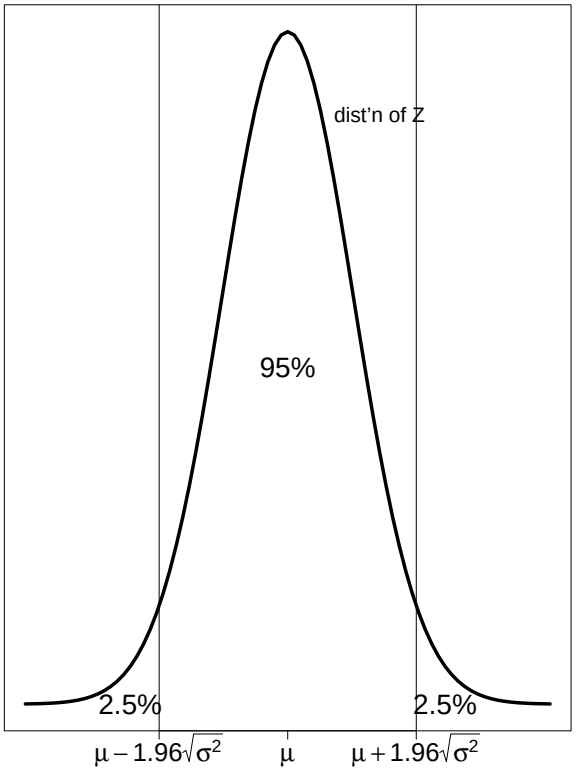
Since $75.6 \neq 0$, can we conclude that H_A is true?

Recall: $b_1 \sim N(\beta_1, \sigma^2/S_{XX})$,
where $S_{XX} = \sum_i (X_i - \bar{X})^2 = 25.38$.

Consider 2 different scenarios:

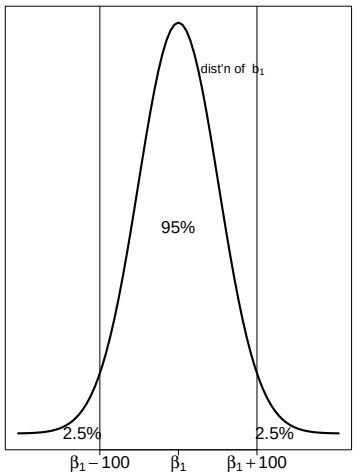
Scenario 1: $\sigma^2/S_{XX} = 2,500 \Rightarrow \sqrt{\sigma^2/S_{XX}} = 50$
Scenario 2: $\sigma^2/S_{XX} = 100 \Rightarrow \sqrt{\sigma^2/S_{XX}} = 10$

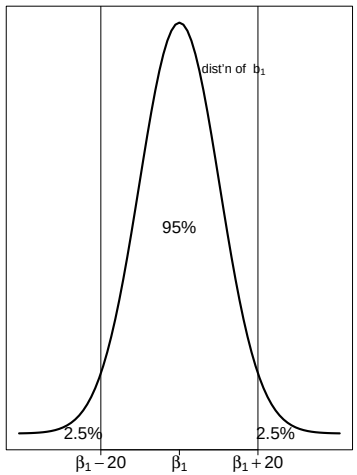
Remember, if $Z \sim N(\mu, \sigma^2)$, then



For the 2 scenarios we have:

Scenario 1: $\sqrt{\sigma^2/S_{XX}} = 50$ Scenario 2: $\sqrt{\sigma^2/S_{XX}} = 10$





Scenario 1: If $\beta_1 = 0$ (H_0 true) then there is a 95% chance that b_1 falls between -100 and 100.
$b_1 = 75.6$ is consistent with $H_0 : \beta_1 = 0$

Scenario 2: If $\beta_1 = 0$ (H_0 true) then there is a 95% chance that b_1 falls between -20 and 20.
$b_1 = 75.6$ suggests that $H_0 : \beta_1 = 0$ is false.

Conclusion: if we know $\sqrt{\sigma^2/S_{XX}}$, we know how likely the value $b_1 = 75.6$ is under H_0, and we can decide if $b_1 = 75.6$ is more consistent with $H_0 : \beta_1 = 0$ or $H_A : \beta_1 \neq 0$.

Last time we showed that

$$b_1 \sim N(\beta_1, \sigma^2/S_{XX}) \quad \Rightarrow \quad \frac{b_1 - \beta_1}{\sqrt{\sigma^2/S_{XX}}} \sim N(0,1)$$

That means that

$$P\left(-1.96 \leq \frac{b_1 - \beta_1}{\sqrt{\sigma^2/S_{XX}}} \leq 1.96\right) = 0.95$$

$$P\left(b_1 - 1.96\sqrt{\frac{\sigma^2}{S_{XX}}} \leq \beta_1 \leq b_1 + 1.96\sqrt{\frac{\sigma^2}{S_{XX}}}\right) = 0.95$$

So, a **95% confidence interval** for β_1 is

$$b_1 \pm 1.96\sqrt{\frac{\sigma^2}{S_{XX}}}$$

Is this a useful confidence interval ? **NO!**

We have to estimate σ^2 under the SLR model.

Remember, the mean squared error

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

is an unbiased estimate of σ^2. So we have all we need!

What's next?

1. tests and confidence intervals for β_1
2. confidence intervals for the mean of Y at some value of X, say X^*, that is, for

$$\beta_0 + \beta_1 X^*$$

3. prediction intervals for the next random variable observed with $X = X^*$

Confidence Intervals and Tests for β_1

The key is $b_1 \sim N(\beta_1, \sigma^2/S_{XX})$. Thus

$$\frac{b_1 - \beta_1}{\sqrt{\sigma^2/S_{XX}}} \sim N(0,1)$$

But this is not useful because we don't know σ^2.

If we replace σ^2 with our estimate of σ^2, MSE, we get

$$\frac{b_1 - \beta_1}{\sqrt{MSE/S_{XX}}} \sim t(n-2).$$

Everything is based on this!

In what follows, α is:
- the type 1 error probability = P(reject H_0 | H_0 true)
- always between 0 and 1 (it's a probability)
- usually set at 0.01, 0.05 or 0.10

$(1-\alpha)100\%$ Confidence Interval for β_1

With probability $1-\alpha$

$$-t(1-\alpha/2; n-2) \leq \frac{b_1 - \beta_1}{\sqrt{MSE/S_{XX}}} \leq t(1-\alpha/2; n-2)$$

Thus, the $(1-\alpha) * 100\%$ confidence interval for β_1 is

$$b_1 \pm t(1-\alpha/2; n-2)\sqrt{MSE/S_{XX}}$$

Don't confuse $t(n-2)$ with $t(1-\alpha/2; n-2)$:

• $t(n-2)$: denotes the type of distribution (t) and its parameter $(n-2)$

• $t(1-\alpha/2; n-2)$: denotes the $1-\alpha/2$ percentile of the $t(n-2)$ distribution (see Table A.2 in the textbook)

α Level Hypothesis Tests concerning β_1

A Two-Sided Test $H_0 : \beta_1 = c$
$H_A : \beta_1 \neq c$

B One-Sided Test $H_0 : \beta_1 \geq c$
$H_A : \beta_1 < c$

C One-Sided Test $H_0 : \beta_1 \leq c$
$H_A : \beta_1 > c$

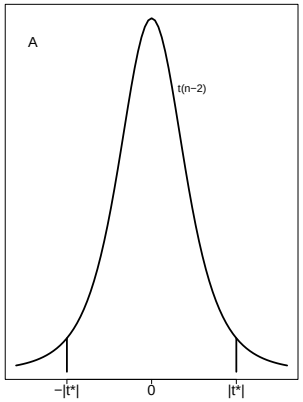
Test Statistic:

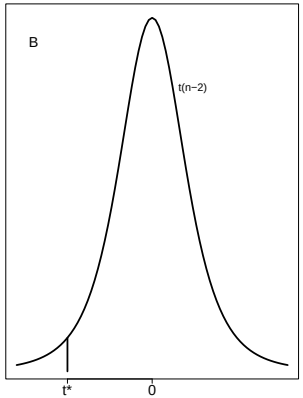
$$t^* = \frac{b_1 - c}{\sqrt{MSE/S_{XX}}}$$

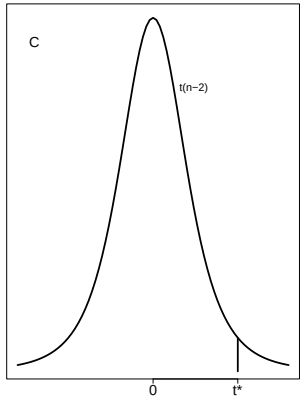
Rejection Rules:

A: reject H_0 if $|t^*| > t(1-\alpha/2; n-2)$
B: reject H_0 if $t^* \leq -t(1-\alpha; n-2)$
C: reject H_0 if $t^* > t(1-\alpha; n-2)$

P-Value: This is the probability of a *more extreme* t^* value than the one we got, given that H_0 is true.







Example of how to do Hypothesis Tests on Homework and Exams:

Question: Test $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ at level $\alpha = 0.05$ for the house prices data. What is the p-value?

$b_1 = 75.6$, $S_{XX} = 25.38$, $MSE = 379.21$

If H_0 is true, then there is no linear relationship between mean house price, $E(Y)$, and square footage.

Answer: $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$, $\alpha = 0.05$

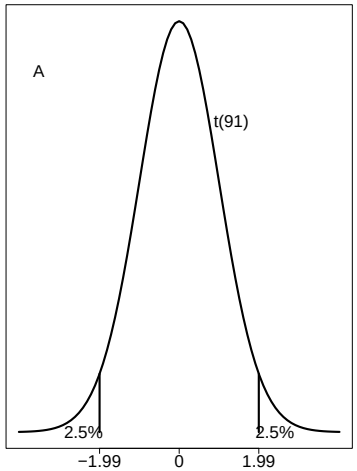
Test Statistic:

$$t^* = \frac{b_1 - 0}{\sqrt{MSE/S_{XX}}} = \frac{75.6}{\sqrt{379.21/25.38}} = 19.56$$

Rejection Rule: Reject H_0 if
$|t^*| > t(1 - \alpha/2; n - 2) = t(0.975; 91) = 1.99.$

Conclusion: Reject H_0 since $19.56 = |t^*| > t(0.975; 91) = 1.99$. There is a significant linear relationship between mean house price and square footage.

Example cont'ed: What's the picture?



Reconsider rejection rule:

$$\begin{aligned} P(\text{reject } H_0 | H_0 \text{ true}) &= P(|t^*| > 1.99 | H_0 \text{ true}) \\ &= 1 - 0.95 = \alpha \end{aligned}$$

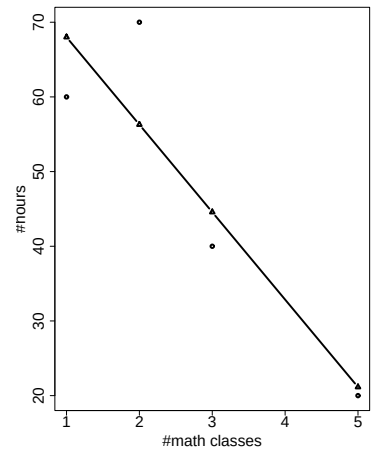
Where is t^* on this picture?
I would have rejected H_0 **for any** $|t^*| > 1.99$!

P-Value: Prob of a more extreme t^* is almost 0.

Extrapolation is Bad!

Never use estimated regression function $\widehat{E}(Y) = b_0 + b_1 X$ outside the range of X values in the data!

Remember the math class/hours on papers example



My friend is taking 7 math classes next semester. How many hours will he spend writing papers?

$80 - 11.7(7) = -1.9 \quad \Rightarrow$ Nice concept, but wrong!

Confidence Intervals for Mean Response

Let X_h denote the level of X for which we wish to estimate the mean response $E(Y_h) = \beta_0 + \beta_1 X_h$.
X_h may be a value which occurred in the sample, or some other value within the scope of the model.

Point estimator $\hat{Y}_h$ of $E(Y_h)$ is

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

Hence, since $b_0 = \sum_i l_i Y_i$ and $b_1 = \sum_i k_i Y_i$ we get

$$\hat{Y}_h = \sum_{i=1}^{n} l_i Y_i + X_h \sum_{i=1}^{n} k_i Y_i = \sum_{i=1}^{n} \Big(l_i + X_h k_i\Big) Y_i$$

Thus $\hat{Y}_h$ is normally dist'd and we can figure out its mean and variance:

$$\begin{aligned} E(\hat{Y}_h) &= \beta_0 + \beta_1 X_h \\ Var(\hat{Y}_h) &= \sigma^2 \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right\} \end{aligned}$$

Together we have

$$\hat{Y}_h \sim N\left(\beta_0 + \beta_1 X_h, \sigma^2 \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right\}\right)$$

or

$$\frac{\hat{Y}_h - (\beta_0 + \beta_1 X_h)}{\sqrt{\sigma^2 \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right\}}} \sim N(0,1)$$

Plug in MSE for the unknown σ^2 gives

$$\frac{\hat{Y}_h - (\beta_0 + \beta_1 X_h)}{\sqrt{MSE \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right\}}} \sim t(n-2)$$

Just like for β_1, a $(1-\alpha)100\%$ CI for $\beta_0 + \beta_1 X_h$ is

$$\hat{Y}_h \pm t(1-\alpha/2; n-2) \sqrt{MSE \left\{ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right\}}$$

Example: Recall for the house data

$$\widehat{E}(price) = -25.2 + 75.6(area)$$

$S_{XX} = 25.38$, $MSE = 379.21$, $\bar{X} = 1.65$

Suppose you are thinking of constructing several 2,000 sq.ft. homes in G'ville and you want to know about how much these houses will sell for.

Point estimate is $\widehat{E}(price) = -25.2 + 75.6(2) = 126$

A 95% CI for $\beta_0 + \beta_1(2)$ is

$$126 \pm t(0.975; 91)\sqrt{379.21\left\{\frac{1}{93} + \frac{(2-1.65)^2}{25.38}\right\}}$$

$= 126 \pm 4.86 \approx (121, 131)$.

Thus, we are 95% confident that the mean selling price of 2,000 sq.ft. houses is between 121,000$ and 131,000$.

(CI for $E(Y_h)$ is smallest for $X_h = \bar{X}$)

Prediction Interval for $Y_{h(new)}$

After we collect the data, we might be interested in predicting a new observation whose X value is X_h.

Before, we estimated the mean of the distribution of Y. Now we predict an individual outcome drawn from the distribution of Y.

Example: There is a 2,000 sq.ft. house about to be put up for sale. Its price is a r.v. $Y_{h(new)}$ and $X_h = 2$.

Suppose that β_0 and β_1 are known.

Question: What do we expect $Y_{h(new)}$ to be?
Answer: $Y_{h(new)} = \beta_0 + \beta_1 X_h + \epsilon_{h(new)}$

So $E(Y_{h(new)}) = \beta_0 + \beta_1 X_h$, $Var(Y_{h(new)}) = \sigma^2$ and

$$Y_{h(new)} \sim N(\beta_0 + \beta_1 X_h, \sigma^2)$$

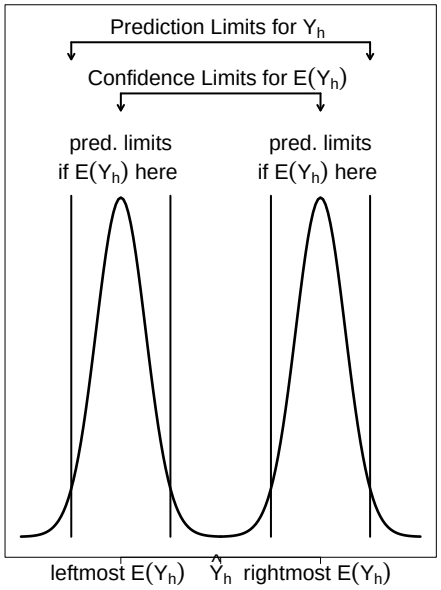
Thus the $1-\alpha$ prediction limits for $Y_{h(new)}$ are:

$$E(Y_{h(new)}) \pm z(1-\alpha/2)\sigma$$

Anyway, we don't know the parameters. But we have a $(1-\alpha)*100\%$ CI for $\beta_0+\beta_1 X_h$:

$$(b_0+b_1X_h)\pm t(1-\alpha/2;n-2)\sqrt{MSE\left\{\frac{1}{n}+\frac{(X_h-\bar{X})^2}{S_{XX}}\right\}}$$

Dist'ns of $Y_{h(new)}$ at the upper and lower CI limit.



The $(1-\alpha)*100\%$ Prediction Interval for $Y_{h(new)}$ is slightly wider than the $(1-\alpha)*100\%$ CI for $\beta_0+\beta_1 X_h$.

We consider the difference

$$Y_{h(new)}-\hat{Y}_h = Y_{h(new)} - \sum_{i=1}^{n}(l_i+X_hk_i)Y_i$$

where $\hat{Y}_h = b_0+b_1X_h$ is indep. of $Y_{h(new)}$. Because it's a linear combination, it's a normal variate with

$$E(Y_{h(new)}-\hat{Y}_h) = E(Y_{h(new)}) - E(\hat{Y}_h) = 0$$

and

$$\begin{aligned} Var(Y_{h(new)}-\hat{Y}_h) &= Var(Y_{h(new)}) + Var(\hat{Y}_h) \\ &= \sigma^2 + \sigma^2\left\{\frac{1}{n}+\frac{(X_h-\bar{X})^2}{S_{XX}}\right\} \\ &= \sigma^2\left\{1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{S_{XX}}\right\} \end{aligned}$$

Thus $(Y_{h(new)} - \hat{Y}_h)/\sqrt{Var(Y_{h(new)} - \hat{Y}_h)} \sim N(0,1)$

$$\frac{Y_{h(new)} - \hat{Y}_h}{\sqrt{MSE\left\{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\right\}}} \sim t(n-2)$$

and a $(1-\alpha) * 100\%$ PI for $Y_{h(new)}$ is given by:

$$\hat{Y}_h \pm t(1-\alpha/2; n-2)\sqrt{MSE\left\{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\right\}}$$

Example: A 95% Prediction Interval for $Y_{h(new)}$, the price of the 2,000 sq.ft. house is

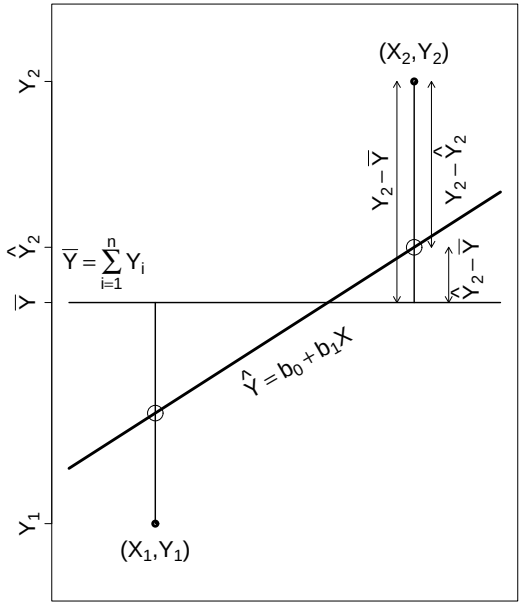
$$126 \pm t(0.975; 91)\sqrt{379.21\left\{1 + \frac{1}{93} + \frac{(2-1.65)^2}{25.38}\right\}}$$

$= 126 \pm 38.5 \approx (87.5, 164.5)$.

Thus, there is a 95% probability that the price of the house will be between 87,500$ and 164,500$.

ANalysis Of VAriance: ANOVA

Nothing new, just a different way of looking at what we have already done.
Say we have the LS estimates of β_0, β_1



Consider the linear relationship

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

Is there a quadratic analogue?

Total Sum of Squares: the variation in the Y's ignoring X

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Regression Sum of Squares: the variation in Y's explained by X

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

Error Sum of Squares: the variation in Y's around the regression line (i.e. not explained by X)

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Does the partition $SSTO = SSR + SSE$ hold? **Yes!**

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Generally, in ANOVA methods, the $SSTO$ is partitioned into several sums of squares which each have an associated **degrees of freedom (df)**.

ANOVA Table for SLR:

Source variat.	Sum of Squares (SS)	df	mean SS
Regr.	$SSR = \sum_i (\hat{Y}_i - \bar{Y})^2$	1	$\frac{SSR}{1}$
Error	$SSE = \sum_i (Y_i - \hat{Y}_i)^2$	$n-2$	$\frac{SSE}{n-2}$
Total	$SSTO = \sum_i (Y_i - \bar{Y})^2$	$n-1$	

Another way to test $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

Test statistic:

$$F^* = \frac{MSR}{MSE}$$

Rejection rule: reject H_0 if $F^* > F(1-\alpha; 1, n-2)$

Fact: F-test and t-test are equivalent; that is the F-test rejects if and only if the t-test rejects.

Notice: using $b_0 = \bar{Y} - b_1\bar{X}$ results in

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n}(b_0 + b_1 X_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(\bar{Y} - b_1\bar{X} + b_1 X_i - \bar{Y})^2 \\ &= b_1^2 \sum_{i=1}^{n}(-\bar{X} + X_i)^2 = b_1^2 S_{XX} \end{aligned}$$

Thus

$$\begin{aligned} F^* = \frac{b_1^2 S_{XX}}{MSE} &= \frac{b_1^2}{MSE/S_{XX}} = \left(\frac{b_1}{\sqrt{MSE/S_{XX}}}\right)^2 \\ &= (t^*)^2 \end{aligned}$$

Generally, if $T \sim t(n-2)$ then $T^2 \sim F(1, n-2)$

Coefficient of Determination, r^2

Question: How strong is the **linear** relationship between Y and X?

Remember: $SSTO = SSR + SSE$

Define:

$$r^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO}$$

with $0 \le r^2 \le 1$.

The higher the r^2, the stronger the linear relationship!

Extreme cases:

- $\hat{Y}_i = Y_i$: then $SSE = 0 \Rightarrow r^2 = 1$
- $b_1 = 0 \Rightarrow \hat{Y}_i = \bar{Y}$: then $SSR = 0 \Rightarrow r^2 = 0$

BUT: $r^2 \approx 0$ does not always mean that there is **no** relationship at all between Y and X! It only means that the relationship is **not linear**!