STAC67: Regression Analysis

Lecture 6

Thibault Randrianarisoa

September 25, 2026

University of Toronto Scarborough

Announcement: Graduate School Open House

Statistical Sciences, University of Toronto

Tomorrow, Saturday 26 September
11am to 3pm
700 University Ave., 9th floor

- Overviews of the MSc in Applied Computing, the Master of Financial Insurance, and the MSc and PhD in Statistics
- Faculty and alumni panels, Q&A, networking lunch

Worth a visit if you are thinking about graduate studies.

Announcement: Quiz 1 next week

In your tutorial: **TUT0001** Tuesday 29 September, 5pm, IA 3120, and **TUT0002** Wednesday 30 September, 3pm, IC 208.

Two parts, 50 minutes in total

1. **Written part first**: 25 minutes, 12 points, on paper.
2. **Then the Quercus part**: 25 minutes, 8 points, four multiple-choice questions on your laptop, with an access code given in the room.

- **Aids**: no aid sheet. A non-programmable calculator is allowed for the written part; R is allowed in the Quercus part. **Laptops stay closed** until the papers are collected.
- **Bring a charged laptop, a calculator and a pen.**
- **Covers** the material up to today's lecture. Reminder: The best two of the three quizzes count.

This replaces what you may have heard earlier: the quiz is not fully written.

Review of Lecture 5

Test for the slope. $H_0 : \beta_1 = \beta_1^0 \quad vs \quad H_a : \beta_1 \neq \beta_1^0$ (or one-sided version, with $<$ or $>$)

Review of Lecture 5

Test for the slope. $H_0 : \beta_1 = \beta_1^0 \quad vs \quad H_a : \beta_1 \neq \beta_1^0$ (or one-sided version, with $<$ or $>$)

$$t^* = \frac{\hat{\beta}_1 - \beta_1^0}{SE(\hat{\beta}_1)} \sim \cancel{N(0,1)} \; t_{n-2} \qquad \text{under } H_0$$

Review of Lecture 5

Test for the slope. $H_0 : \beta_1 = \beta_1^0 \quad vs \quad H_a : \beta_1 \neq \beta_1^0$ (or one-sided version, with $<$ or $>$)

$$t^* = \frac{\hat{\beta}_1 - \beta_1^0}{SE(\hat{\beta}_1)} \sim \cancel{N(0,1)} \; t_{n-2} \qquad \text{under } H_0$$

Not normal, because σ is estimated, not known:

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}, \qquad SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

Review of Lecture 5

Test for the slope. $H_0 : \beta_1 = \beta_1^0 \quad vs \quad H_a : \beta_1 \neq \beta_1^0$ (or one-sided version, with $<$ or $>$)

$$t^* = \frac{\hat{\beta}_1 - \beta_1^0}{SE(\hat{\beta}_1)} \sim \cancel{N(0,1)} \; t_{n-2} \qquad \text{under } H_0$$

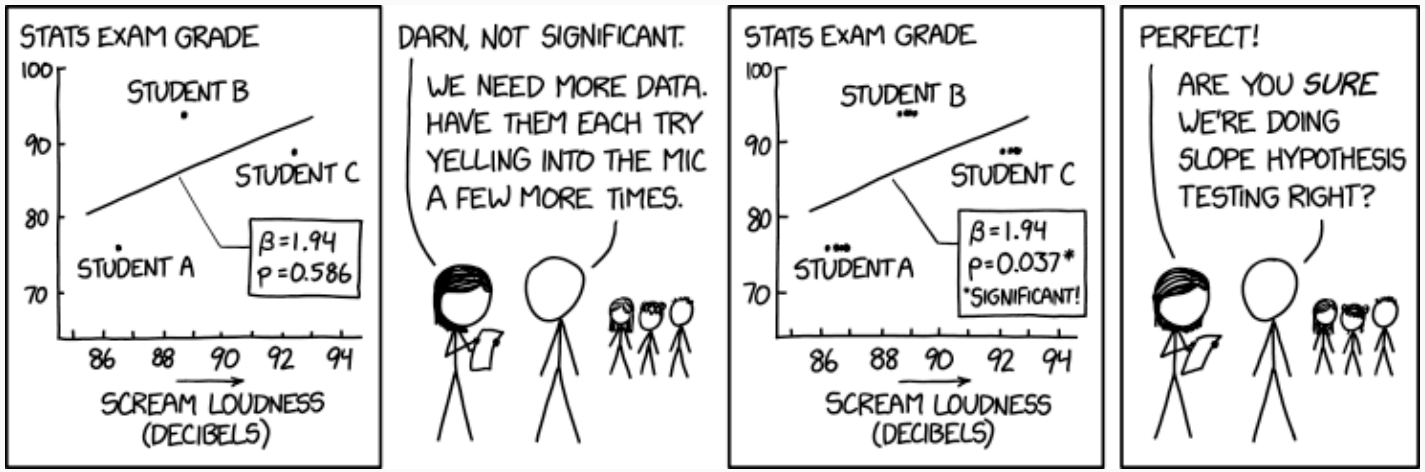
Not normal, because σ is estimated, not known:

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}, \qquad SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

$100(1-\alpha)\%$ confidence interval for β_1: estimate $\pm$ margin of error

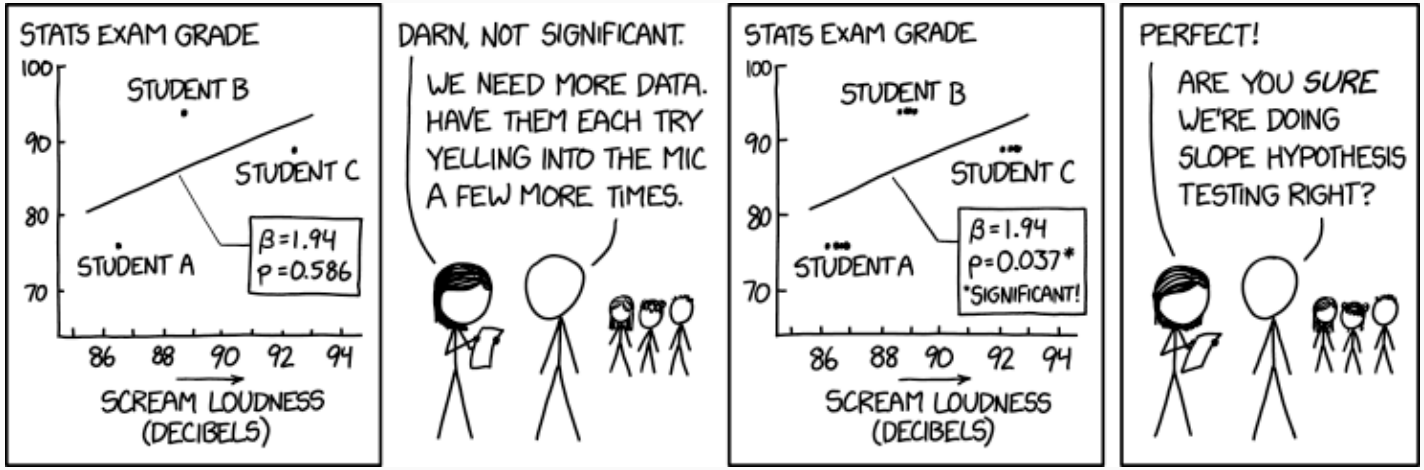
$$\hat{\beta}_1 \pm \underbrace{t(1-\alpha/2;\, n-2)}_{\textbf{student quantile}} \cdot \underbrace{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}}_{SE(\hat{\beta}_1)}$$

Back to the comic: why does the p-value fall?



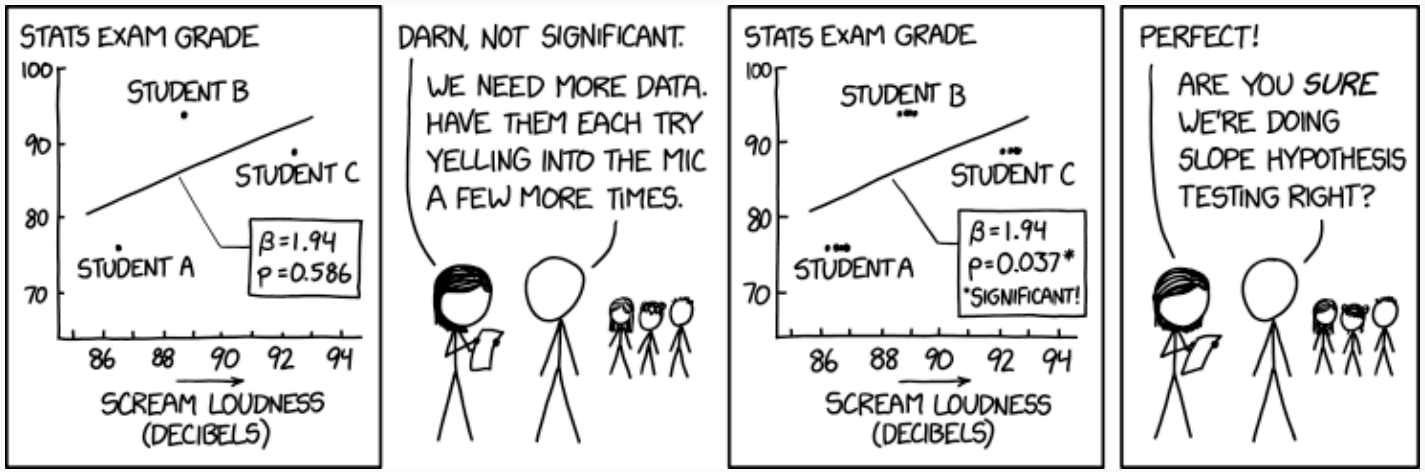
- The new points are the **same three students**, measured again: same grade, same everything.

Back to the comic: why does the p-value fall?



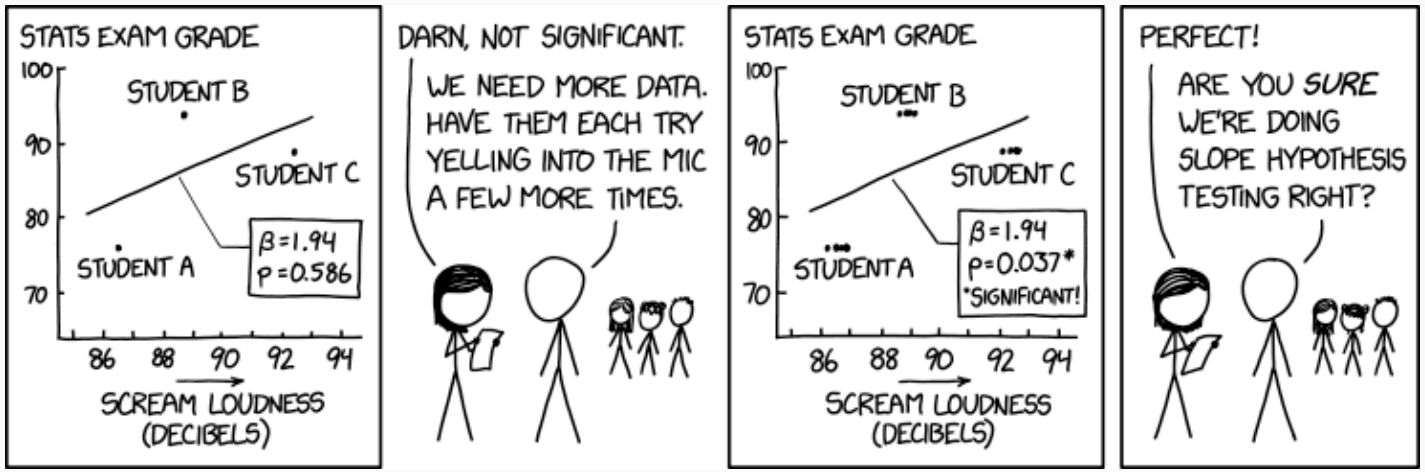
- The new points are the **same three students**, measured again: same grade, same everything.
- Repeating the points does not change the estimate, so $\hat{\beta}_1 = 1.94$ in both panels.

Back to the comic: why does the p-value fall?



- The new points are the **same three students**, measured again: same grade, same everything.
- Repeating the points does not change the estimate, so $\hat{\beta}_1 = 1.94$ in both panels.
- But formulas are based on the assumption that there are 9 *independent* observations, instead of 3:
  - $SE(\hat{\beta}_1) = \hat{\sigma}/\sqrt{S_{xx}}$, and S_{xx} triples, so the standard error shrinks,
  - the degrees of freedom go from $n - 2 = 1$ to 7: $t(0.975;\, 1) = 12.71$, but $t(0.975;\, 7) = 2.36$.

Back to the comic: why does the p-value fall?



- The new points are the **same three students**, measured again: same grade, same everything.
- Repeating the points does not change the estimate, so $\hat{\beta}_1 = 1.94$ in both panels.
- But formulas are based on the assumption that there are 9 *independent* observations, instead of 3:
  - $SE(\hat{\beta}_1) = \hat{\sigma}/\sqrt{S_{xx}}$, and S_{xx} triples, so the standard error shrinks,
  - the degrees of freedom go from $n - 2 = 1$ to 7: $t(0.975;\, 1) = 12.71$, but $t(0.975;\, 7) = 2.36$.

⚠ The **independence** assumption fails: measuring the same students again is not new information, so the p-value is not to be trusted.

Exercise (Crime Rate)

Let's go back to the example from last week.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20517.5999	3277.64269	6.259865	0.00e+00
X	-170.5752	41.57433	-4.102897	9.57e-05

We want to test whether there is a linear association between crime rate and the percentage of high school graduates, using a t test with $\alpha = 0.01$. State the hypotheses, decision rule, and conclusion.

Exercise (Crime Rate)

Let's go back to the example from last week.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20517.5999	3277.64269	6.259865	0.00e+00
X	-170.5752	41.57433	-4.102897	9.57e-05

We want to test whether there is a linear association between crime rate and the percentage of high school graduates, using a t test with $\alpha = 0.01$. State the hypotheses, decision rule, and conclusion.

- Hypotheses: $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$.

Exercise (Crime Rate)

Let's go back to the example from last week.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20517.5999	3277.64269	6.259865	0.00e+00
X	-170.5752	41.57433	-4.102897	9.57e-05

We want to test whether there is a linear association between crime rate and the percentage of high school graduates, using a t test with $\alpha = 0.01$. State the hypotheses, decision rule, and conclusion.

- Hypotheses: $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$.
- Test statistic: $t^* = \dfrac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \dfrac{-170.575}{41.574} = -4.103$, on $n - 2 = 82$ degrees of freedom.

Exercise (Crime Rate)

Let's go back to the example from last week.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20517.5999	3277.64269	6.259865	0.00e+00
X	-170.5752	41.57433	-4.102897	9.57e-05

We want to test whether there is a linear association between crime rate and the percentage of high school graduates, using a t test with $\alpha = 0.01$. State the hypotheses, decision rule, and conclusion.

- Hypotheses: $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$.
- Test statistic: $t^* = \dfrac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \dfrac{-170.575}{41.574} = -4.103,$ on $n - 2 = 82$ degrees of freedom.
- Decision rule: reject H_0 if $|t^*| > t(0.995;\, 82) = 2.637$. Here $|t^*| = 4.103 > 2.637$, so we **reject** H_0. Equivalently, p-value $= 0.0001 < 0.01$.

Exercise (Crime Rate)

Let's go back to the example from last week.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20517.5999	3277.64269	6.259865	0.00e+00
X	-170.5752	41.57433	-4.102897	9.57e-05

We want to test whether there is a linear association between crime rate and the percentage of high school graduates, using a t test with $\alpha = 0.01$. State the hypotheses, decision rule, and conclusion.

- Hypotheses: $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$.
- Test statistic: $t^* = \dfrac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \dfrac{-170.575}{41.574} = -4.103,$ on $n - 2 = 82$ degrees of freedom.
- Decision rule: reject H_0 if $|t^*| > t(0.995;\, 82) = 2.637$. Here $|t^*| = 4.103 > 2.637$, so we **reject** H_0. Equivalently, p-value $= 0.0001 < 0.01$.
- **Conclusion**: at the 1% level, there is evidence of a linear association between crime rate and the percentage of high school graduates.

Exercise (Crime Rate)

Estimate β_1 with a 99 % confidence interval. Interpret the interval estimate.

Exercise (Crime Rate)

Estimate β_1 with a 99 % confidence interval. Interpret the interval estimate.

- $n = 84$, so $n - 2 = 82$ degrees of freedom, and $t(0.995;\, 82) = 2.637$.

$$\begin{aligned} \hat{\beta}_1 \pm t(0.995;\, 82)\, SE(\hat{\beta}_1) &= -170.575 \pm 2.637 \times 41.574 \\ &= (-280.21,\ -60.94) \end{aligned}$$

Exercise (Crime Rate)

Estimate β_1 with a 99 % confidence interval. Interpret the interval estimate.

- $n = 84$, so $n - 2 = 82$ degrees of freedom, and $t(0.995;\, 82) = 2.637$.

$$\begin{aligned} \hat{\beta}_1 \pm t(0.995;\, 82)\, SE(\hat{\beta}_1) &= -170.575 \pm 2.637 \times 41.574 \\ &= (-280.21,\ -60.94) \end{aligned}$$

- **Interpretation** With 99% confidence, the mean crime rate falls by between 60.9 and 280.2 offences per 100,000 for each additional percentage point of high-school graduates.

Exercise (Crime Rate)

Estimate β_1 with a 99 % confidence interval. Interpret the interval estimate.

- $n = 84$, so $n - 2 = 82$ degrees of freedom, and $t(0.995;\, 82) = 2.637$.

$$\begin{aligned} \hat{\beta}_1 \pm t(0.995;\, 82)\, SE(\hat{\beta}_1) &= -170.575 \pm 2.637 \times 41.574 \\ &= (-280.21,\ -60.94) \end{aligned}$$

- **Interpretation** With 99% confidence, the mean crime rate falls by between 60.9 and 280.2 offences per 100,000 for each additional percentage point of high-school graduates.
- Zero is **not** in the interval, so at the 1% level there is evidence of a linear relationship between X and Y.

In R: `confint(fit, level = 0.99)`.

Interval Estimation of $E(Y|X = x_0)$

Suppose that x_{n+1} is a new value of x for which we want to do prediction.

1. Estimation of $\mu_{n+1} = E[Y|X = x_{n+1}]$
2. Prediction of Y value for an individual with $X = x_{n+1}$
   - We use fitted regression model to do both of these
   - Estimation of

$$\mu_{n+1} = \beta_0 + \beta_1 X_{n+1}$$

Interval Estimation of $E(Y|X = x_0)$

Suppose that x_{n+1} is a new value of x for which we want to do prediction.

1. Estimation of $\mu_{n+1} = E[Y|X = x_{n+1}]$
2. Prediction of Y value for an individual with $X = x_{n+1}$

- We use fitted regression model to do both of these
- Estimation of

$$\mu_{n+1} = \beta_0 + \beta_1 X_{n+1}$$

- We estimate it by the fitted value at x_{n+1}:

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}, \qquad E(\hat{Y}_{n+1}) = \beta_0 + \beta_1 x_{n+1} = \mu_{n+1} \text{ (unbiased)}$$

Interval Estimation of $E(Y|X = x_0)$

Suppose that x_{n+1} is a new value of x for which we want to do prediction.

1. Estimation of $\mu_{n+1} = E[Y|X = x_{n+1}]$
2. Prediction of Y value for an individual with $X = x_{n+1}$

- We use fitted regression model to do both of these
- Estimation of

$$\mu_{n+1} = \beta_0 + \beta_1 X_{n+1}$$

- We estimate it by the fitted value at x_{n+1}:

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}, \qquad E(\hat{Y}_{n+1}) = \beta_0 + \beta_1 x_{n+1} = \mu_{n+1} \text{ (unbiased)}$$

Let's now derive the variance formula as well.

Derivation of $Var(\hat{Y}_{n+1})$

Recall from Lecture 4, the variances of the estimators:

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Derivation of $Var(\hat{Y}_{n+1})$

Recall from Lecture 4, the variances of the estimators:

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Re-centre the fitted value, using $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$:

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} = \bar{Y} + \hat{\beta}_1 (x_{n+1} - \bar{x})$$

Derivation of $Var(\hat{Y}_{n+1})$

Recall from Lecture 4, the variances of the estimators:

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Re-centre the fitted value, using $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$:

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} = \bar{Y} + \hat{\beta}_1 (x_{n+1} - \bar{x})$$

Also from Lecture 4: We have two terms that are uncorrelated $Cov(\bar{Y}, \hat{\beta}_1) = 0$, so the variances add:

$$Var(\hat{Y}_{n+1}) = Var(\bar{Y}) + (x_{n+1} - \bar{x})^2 Var(\hat{\beta}_1) = \frac{\sigma^2}{n} + (x_{n+1} - \bar{x})^2 \frac{\sigma^2}{S_{xx}}$$

Derivation of $Var(\hat{Y}_{n+1})$

Recall from Lecture 4, the variances of the estimators:

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Re-centre the fitted value, using $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$:

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} = \bar{Y} + \hat{\beta}_1 (x_{n+1} - \bar{x})$$

Also from Lecture 4: We have two terms that are uncorrelated $Cov(\bar{Y}, \hat{\beta}_1) = 0$, so the variances add:

$$Var(\hat{Y}_{n+1}) = Var(\bar{Y}) + (x_{n+1} - \bar{x})^2 Var(\hat{\beta}_1) = \frac{\sigma^2}{n} + (x_{n+1} - \bar{x})^2 \frac{\sigma^2}{S_{xx}}$$

$$\boxed{Var(\hat{Y}_{n+1}) = \sigma^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right]} \qquad SE(\hat{Y}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

Derivation of $Var(\hat{Y}_{n+1})$

Analysis

- It is the smallest at $x_{n+1} = \bar{x}$, and growing as you move away: the line is best determined *in the middle of the data*.
- That is why, as we will see below, the confidence band is pinched at the centre and flares at the ends.
- Also, it decreases as n or the variability of X grows.

Sanity check: at $x_{n+1} = 0$, $\hat{Y}_{n+1} = \hat{\beta}_0$, and the box gives back $Var(\hat{\beta}_0)$ above.

Confidence Interval for $\mu_{n+1} = E(Y|X = x_{n+1})$

- A $(1-\alpha)\%$ confidence interval for μ_{n+1}:

$$\hat{Y}_{n+1} \pm t(1-\alpha/2; n-2)SE(\hat{Y}_{n+1})$$

- Exercise: Obtain 95% confidence interval for the mean crime rate for states of high school graduate rate of 80%.

```
new.data = data.frame(X=80)
predict(fit, new.data, interval="confidence")
```

```
##        fit      lwr      upr
## 1 6871.585 6347.116 7396.054
```

Prediction of new observation

The value of Y for an individual with $X = x_{n+1}$ is:

$$Y_{n+1} = \beta_0 + \beta_1 X + \epsilon_0$$

- Estimate of Y_{n+1}: $\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1} = \hat{\mu}_0$
- Prediction error: $e_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$
- $Var(e_{n+1}) = Var(Y_{n+1} - \hat{Y}_{n+1}) = Var(Y_{n+1}) + Var(\hat{Y}_{n+1})$, since the new observation (x_{n+1}, Y_{n+1}) is not correlated with the sample the line was fitted on

  $$= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right]$$

  $$s_{\{pred\}} = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

  The extra 1 is the noise of the new observation itself. It never shrinks, so a prediction interval stays wide no matter how much data you collect.
- $100 \times (1 - \alpha)\%$ prediction Interval:

$$\hat{Y}_{n+1} \pm t(1 - \alpha/2; n - 2) s_{\{pred\}}$$

Exercise

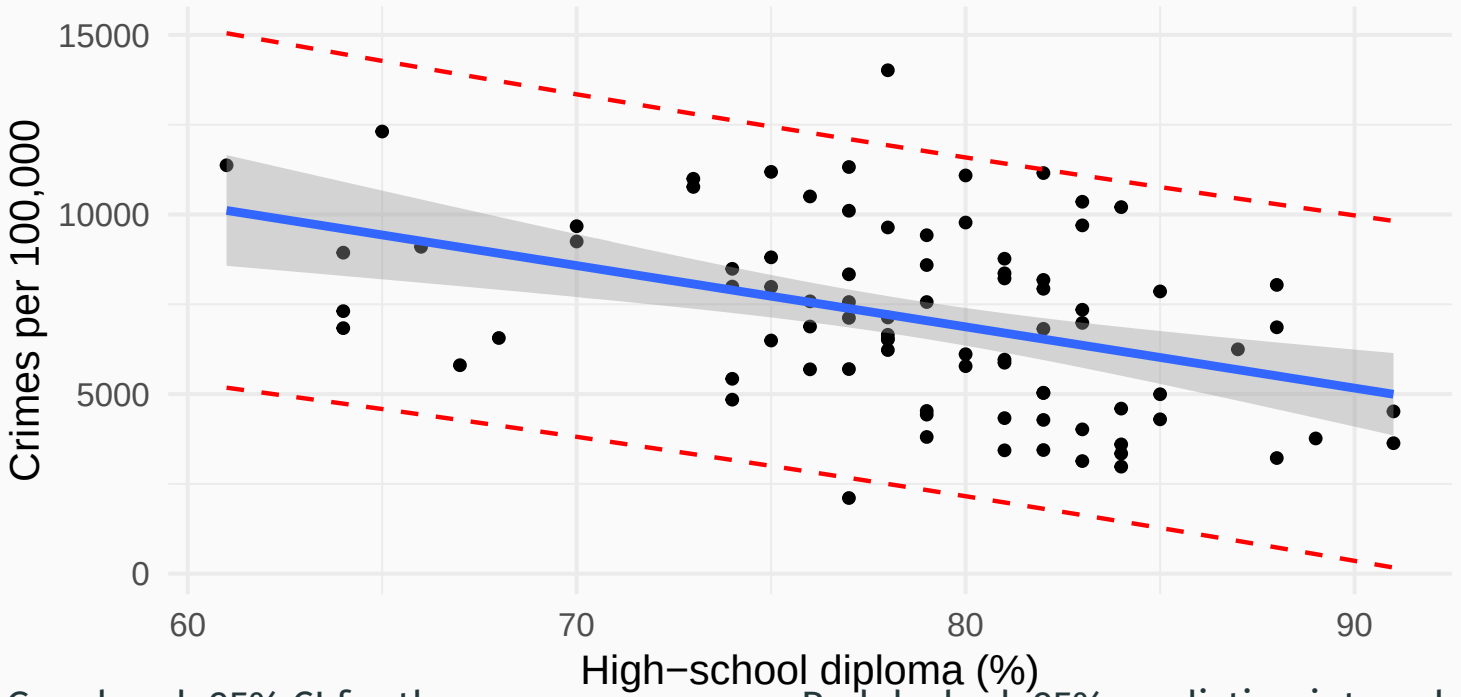
- Exercise: Obtain 95% prediction interval for the crime rate for states of high school graduate rate of 80%.

```
new.data = data.frame(X=80)
predict(fit, new.data, interval="prediction")
```

```
##        fit     lwr      upr
## 1 6871.585 2154.92 11588.25
```

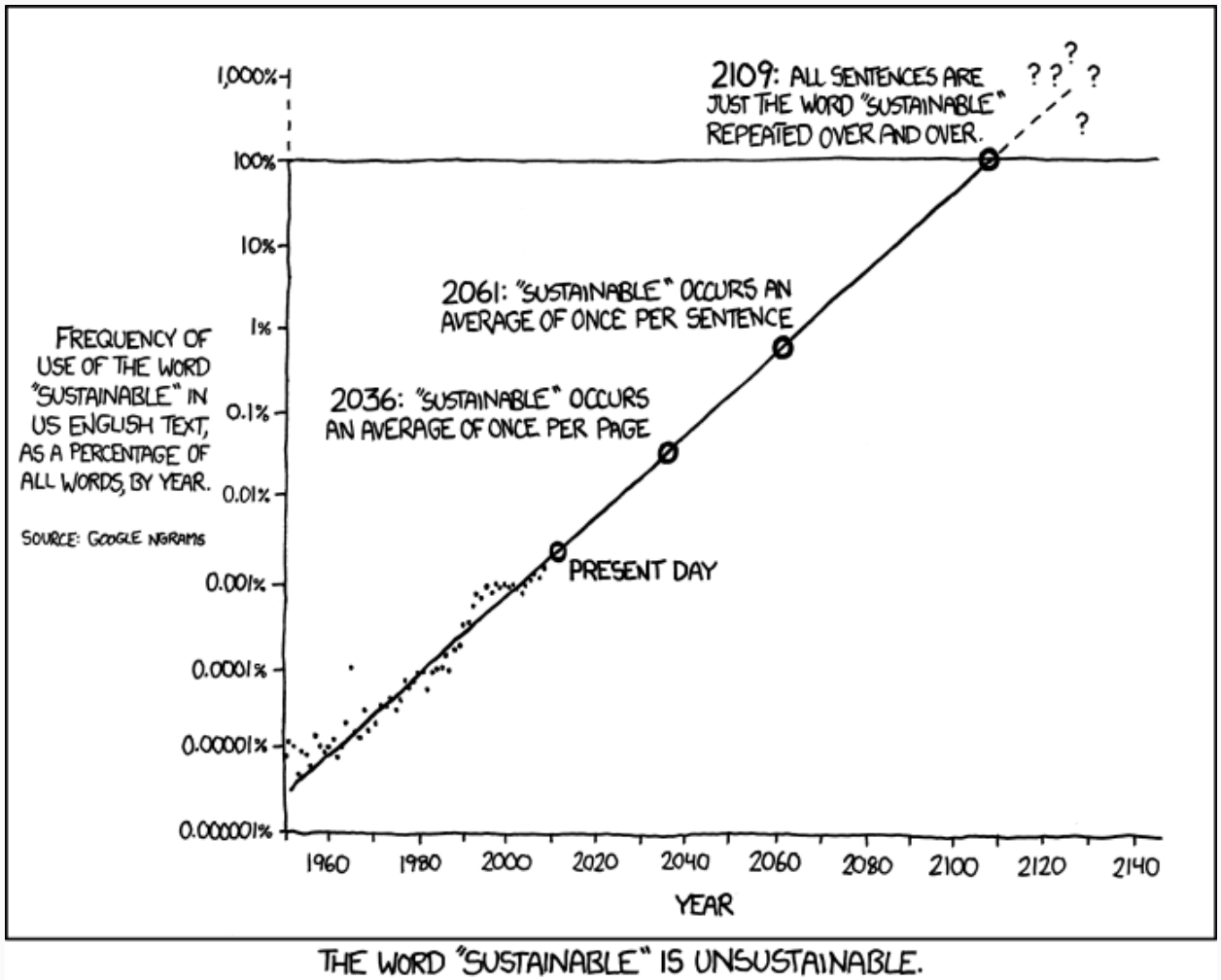
Graph Prediction Intervals (using ggplot)

```r
library(ggplot2)

Crime.Pred <- predict(fit, interval = "prediction")
new.df <- cbind(Crime, Crime.Pred)

ggplot(new.df, aes(X, Y)) +
  geom_point(size = 0.9) +
  geom_line(aes(y = lwr), colour = "red", linetype = "dashed") +
  geom_line(aes(y = upr), colour = "red", linetype = "dashed") +
  geom_smooth(method = lm, se = TRUE) +
  labs(x = "High-school diploma (%)", y = "Crimes per 100,000") +
  theme_minimal(base_size = 10)
```



Grey band: 95% CI for the mean response. Red dashed: 95% prediction interval.

How far can a line be pushed?



A line fitted to sixty years of data, extended a century past them. The $(x_0 - \bar{x})^2$ term widens our intervals away from the data, but only on the assumption that the line still holds out there. No interval protects against the model itself breaking.

Analysis of Variance Approach to Regression Analysis

- How well does the least squares fit explain variation in Y?

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

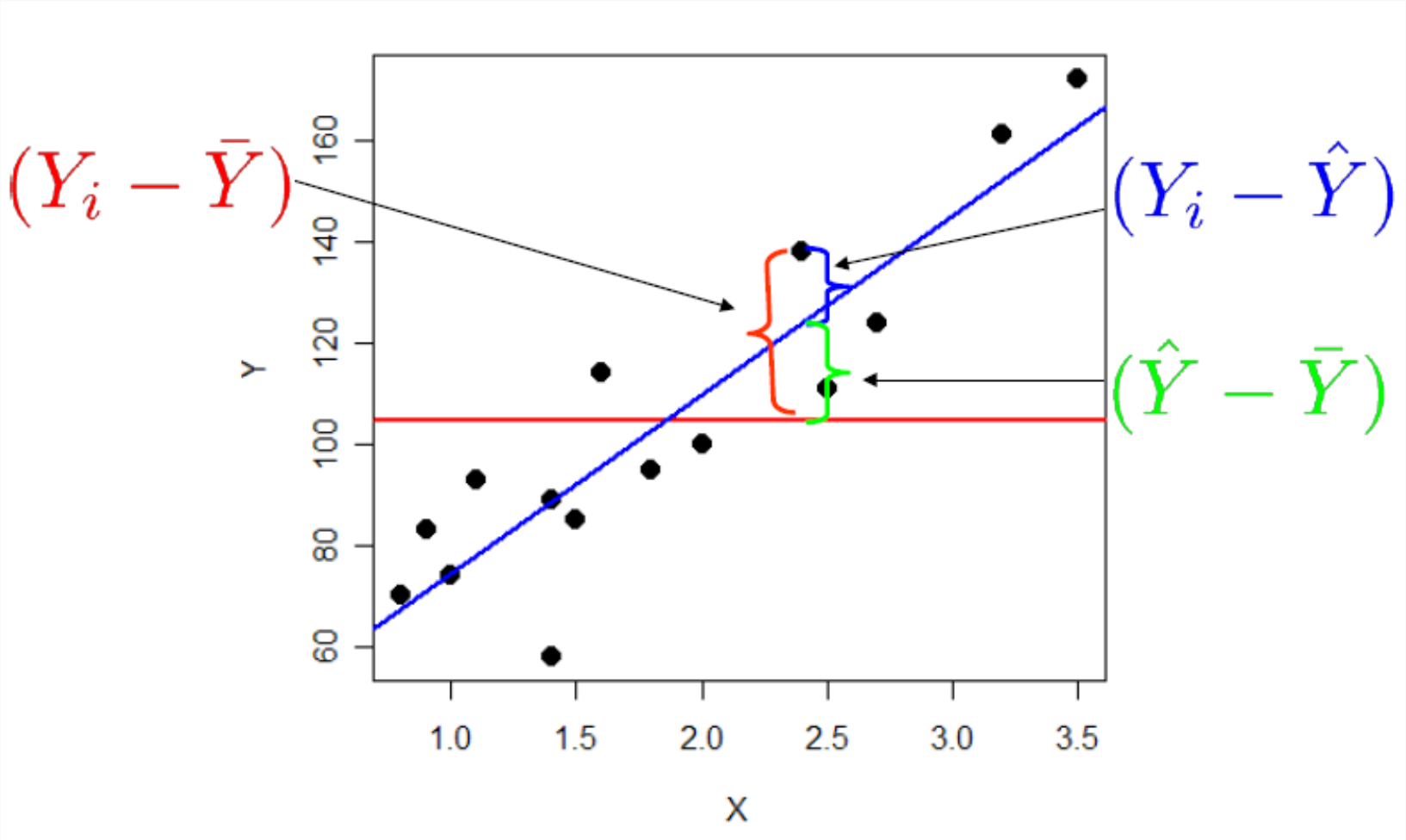
$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

- Total Sum of Squares (SST):

- Model Sum of Squares (SSR): Variation in Y explained by the regression.

- Error Sum of Squares (SSE): Variation in Y that is left explained.

How does that breakdown look on a scatterplot?



Analysis of Variance Table

Source of Variation	Sum of Squares	df	Mean Squares
Regression	SSR	1	$MSR = SSR/1$
Error	SSE	$n-2$	$MSE = SSE/(n-2)$
Total	SST	$n-1$	

- The total degrees of freedom is always $n-1$, as we fit the mean only.
- In the simple regression, the degrees of freedom used by the model is $n-2$: intercept and slope are fitted.
- F-test for $H_0 : \beta_1 = 0$

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/n-2}$$

- Under $H_0, \quad F \sim F(1, n-2)$

F distribution

- The F-distribution with k_1 and k_2 degrees of freedom can be defined as the distribution of the random variable F

$$F = \frac{V_1/k_1}{V_2/k_2},$$

F distribution

- The F-distribution with k_1 and k_2 degrees of freedom can be defined as the distribution of the random variable F

$$F = \frac{V_1/k_1}{V_2/k_2},$$

where $V_1 \sim \chi^2_{k_1}$, $V_2 \sim \chi^2_{k_2}$, and V_1 and V_2 are independent.

F distribution

- The F-distribution with k_1 and k_2 degrees of freedom can be defined as the distribution of the random variable F

$$F = \frac{V_1/k_1}{V_2/k_2},$$

  where $V_1 \sim \chi^2_{k_1}, \quad V_2 \sim \chi^2_{k_2}, \quad$ and V_1 and V_2 are independent.
- This is denoted as $F \sim F(k_1, k_2)$
- we can show that under $H_0 : \beta_1 = 0$,

$$\frac{MSR}{MSE} \sim F(1, n-2)$$

Relationship b/w F-test and t-test

- We can rewrite SSR using the regression estimator:

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} \left(\hat{\beta}_1 (x_i - \bar{x})\right)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 \, S_{xx}$$

Relationship b/w F-test and t-test

- We can rewrite SSR using the regression estimator:

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} \left(\hat{\beta}_1 (x_i - \bar{x})\right)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 \, S_{xx}$$

using $\hat{Y}_i - \bar{Y} = \hat{\beta}_1 (x_i - \bar{x})$, the same re-centring as on the $Var(\hat{Y}_0)$ slide.

Relationship b/w F-test and t-test

- We can rewrite SSR using the regression estimator:

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} \left(\hat{\beta}_1 (x_i - \bar{x})\right)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 \, S_{xx}$$

using $\hat{Y}_i - \bar{Y} = \hat{\beta}_1 (x_i - \bar{x})$, the same re-centring as on the $Var(\hat{Y}_0)$ slide.

$$F^* = \frac{SSR/1}{SSE/(n-2)} = \frac{\hat{\beta}_1^2 \, S_{xx}}{MSE} = \frac{\hat{\beta}_1^2}{MSE/S_{xx}} = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}\right)^2 = (t^*)^2$$

Relationship b/w F-test and t-test

- We can rewrite SSR using the regression estimator:

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} \left(\hat{\beta}_1 (x_i - \bar{x})\right)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 \, S_{xx}$$

using $\hat{Y}_i - \bar{Y} = \hat{\beta}_1 (x_i - \bar{x})$, the same re-centring as on the $Var(\hat{Y}_0)$ slide.

$$F^* = \frac{SSR/1}{SSE/(n-2)} = \frac{\hat{\beta}_1^2 \, S_{xx}}{MSE} = \frac{\hat{\beta}_1^2}{MSE/S_{xx}} = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}\right)^2 = (t^*)^2$$

- In the simple regression, this is equivalent to the test of

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

Descriptive Measure of Linear Association b/w X and Y

$$SST = SSR + SSE$$

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

$\frac{SSR}{SST}$: the proportion of the total variation in Y explained by the regression

- A "good" model should have a large $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- R^2: Coefficient of determination

Coefficient of Determination

Properties

1. $0 \leq R^2 \leq 1$
2. In simple linear regression, $R^2 = r^2$

Coefficient of Determination

Properties

1. $0 \leq R^2 \leq 1$
2. In simple linear regression, $R^2 = r^2$

Indeed:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = S_{YY}, \qquad SSR = \hat{\beta}_1^2 \, S_{xx} \text{ (slide 20)}$$

Coefficient of Determination

Properties

1. $0 \leq R^2 \leq 1$
2. In simple linear regression, $R^2 = r^2$

Indeed:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = S_{YY}, \qquad SSR = \hat{\beta}_1^2 \, S_{xx} \text{ (slide 20)}$$

and, from Lecture 3, $\hat{\beta}_1 = r\sqrt{S_{YY}/S_{xx}}$. Therefore

$$R^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1^2 \, S_{xx}}{S_{YY}} = r^2 \, \frac{S_{YY}}{S_{xx}} \cdot \frac{S_{xx}}{S_{YY}} = r^2$$

Coefficient of Determination

Properties

1. $0 \leq R^2 \leq 1$
2. In simple linear regression, $R^2 = r^2$

Indeed:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = S_{YY}, \qquad SSR = \hat{\beta}_1^2 \, S_{xx} \text{ (slide 20)}$$

and, from Lecture 3, $\hat{\beta}_1 = r\sqrt{S_{YY}/S_{xx}}$. Therefore

$$R^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1^2 \, S_{xx}}{S_{YY}} = r^2 \, \frac{S_{YY}}{S_{xx}} \cdot \frac{S_{xx}}{S_{YY}} = r^2$$

Only in simple linear regression: we will see later that with several predictors, R^2 is not the square of a correlation.

The F statistic under H_0

- Show that $\frac{MSR}{MSE} \sim F(1, n-2)$.

The F statistic under H_0

- Show that $\frac{MSR}{MSE} \sim F(1, n-2)$.

Under $H_0 : \beta_1 = 0$ we have $\hat{\beta}_1 \sim N\left(0, \sigma^2/S_{xx}\right)$, so $\hat{\beta}_1\sqrt{S_{xx}}/\sigma \sim N(0,1)$ and therefore

$$\frac{SSR}{\sigma^2} = \frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} = \left(\frac{\hat{\beta}_1\sqrt{S_{xx}}}{\sigma}\right)^2 \sim \chi^2_{(1)}$$

The F statistic under H_0

- Show that $\frac{MSR}{MSE} \sim F(1, n-2)$.

Under $H_0 : \beta_1 = 0$ we have $\hat{\beta}_1 \sim N\left(0, \sigma^2/S_{xx}\right)$, so $\hat{\beta}_1\sqrt{S_{xx}}/\sigma \sim N(0,1)$ and therefore

$$\frac{SSR}{\sigma^2} = \frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} = \left(\frac{\hat{\beta}_1\sqrt{S_{xx}}}{\sigma}\right)^2 \sim \chi^2_{(1)}$$

We already know $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-2)}$, and SSR and SSE are independent (admitted).

So

$$\frac{MSR}{MSE} = \frac{(SSR/\sigma^2)/1}{(SSE/\sigma^2)/(n-2)} \sim F(1, n-2)$$

The F statistic under H_0

- Show that $\frac{MSR}{MSE} \sim F(1, n-2)$.

Under $H_0 : \beta_1 = 0$ we have $\hat{\beta}_1 \sim N\left(0, \sigma^2 / S_{xx}\right)$, so $\hat{\beta}_1 \sqrt{S_{xx}}/\sigma \sim N(0,1)$ and therefore

$$\frac{SSR}{\sigma^2} = \frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} = \left(\frac{\hat{\beta}_1 \sqrt{S_{xx}}}{\sigma}\right)^2 \sim \chi^2_{(1)}$$

We already know $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-2)}$, and SSR and SSE are independent (admitted).

So

$$\frac{MSR}{MSE} = \frac{(SSR/\sigma^2)/1}{(SSE/\sigma^2)/(n-2)} \sim F(1, n-2)$$

Two independent chi-squares, each over its degrees of freedom.

Example (Crime Rate)

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df    Sum Sq  Mean Sq F value    Pr(>F)
## X          1  93462942 93462942  16.834 9.571e-05 ***
## Residuals 82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1) Test whether or not there is a linear association between crime rate and percentage of high school graduates using an F test. Show the numerical equivalence of the two test statistics and decision rules.

Example (Crime Rate)

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df    Sum Sq  Mean Sq F value    Pr(>F)
## X          1  93462942 93462942  16.834 9.571e-05 ***
## Residuals 82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1) Test whether or not there is a linear association between crime rate and percentage of high school graduates using an F test. Show the numerical equivalence of the two test statistics and decision rules.

$$F^* = \frac{MSR}{MSE} = \frac{93{,}462{,}942}{5{,}552{,}112} = 16.83 \text{ on } (1,\ 82) \text{ df}, \qquad p = 0.0001 : \textbf{ reject } H_0$$

Example (Crime Rate)

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df    Sum Sq  Mean Sq F value    Pr(>F)
## X          1  93462942 93462942  16.834 9.571e-05 ***
## Residuals 82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1) Test whether or not there is a linear association between crime rate and percentage of high school graduates using an F test. Show the numerical equivalence of the two test statistics and decision rules.

$$F^* = \frac{MSR}{MSE} = \frac{93{,}462{,}942}{5{,}552{,}112} = 16.83 \text{ on } (1,\ 82) \text{ df}, \qquad p = 0.0001 : \textbf{ reject } H_0$$

$$(t^*)^2 = (-4.103)^2 = 16.83 = F^* \qquad \text{the two tests are one}$$

Example (Crime Rate), continued

2) Compute R^2 and r.

Example (Crime Rate), continued

2) Compute R^2 and r.

$$R^2 = \frac{SSR}{SST} = \frac{93{,}462{,}942}{548{,}736{,}108} = 0.170, \qquad r = -\sqrt{0.170} = -0.413$$

Example (Crime Rate), continued

2) Compute R^2 and r.

$$R^2 = \frac{SSR}{SST} = \frac{93{,}462{,}942}{548{,}736{,}108} = 0.170, \qquad r = -\sqrt{0.170} = -0.413$$

⚠ $R^2 = 0.17$ is small, yet the slope is significant at the 1% level.

- Significance answers *is there a relationship?*
- R^2 answers *how much of the variation does it explain?*
- They are different questions, and with $n = 84$ a weak relationship can still be clearly non-zero.