Simple Linear Regression II

Emilija Perković

Dept. of Statistics
University of Washington

Goal of hypothesis testing

In hypothesis testing, our goal is to test whether a parameter estimate is significantly different from some pre-determined value.

Often (but not always), we are interested to see if the estimate(s) are significantly different from zero.

We will look at hypothesis tests for the estimate of the slope. Hypothesis tests for the estimated intercept can be constructed analogously.

Example: Hypothesis testing

Example: Snowfall data `ftcollinssnow` from R package `alr4`.

Can early season (Sept 1 - Dec 31) snowfall predict snowfall for the remainder of the season (Jan 1 - June 30)?

Data: The amounts of snowfall (in inches) for 93 years in Ft. Collins.

Let y denote the amount of late season snowfall, and x denote the amount of early season snowfall.
Given the regression model (and given that ϵ_i iid $\mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$):

$$y = \beta_0 + \beta_1 x + \epsilon$$

the test of interest is:

$$H_0 : \beta_1 = \beta_1^* \text{ (and } H_A : \beta_1 \neq \beta_1^*).$$

Example: Hypothesis testing

Assume ϵ_i iid $\mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$. For testing the null hypothesis:

$$H_0 : \beta_1 = \beta_1^* \text{ (and } H_A : \beta_1 \neq \beta_1^*).$$

we can compute the t-statistic as

$$T = \frac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1 | X = x)}$$

where β_1^* is the value from the null hypothesis.

The t-statistic follows Student's t_{n-2} distribution under the null hypothesis:

$$T | X = x \sim t_{n-2}.$$

Example: Hypothesis testing

We make a test decision based on the p-value. Recall: a p-value is

"The probability, under the null hypothesis, of obtaining a result as or more extreme than the observed result."

If t is the observed value of our test statistic, then the p-value of the test is calculated as

$$P(|T| \geq |t| \,\big|H_0), T \sim t_{n-2}.$$

Example: Hypothesis testing

Let $\beta_1^* = 0$. Given the estimated slope and its standard error for Ft. Collins snowfall data over 93 years

$$\hat{\beta}_1 = 0.2035, \; SE(\hat{\beta}_1|X = x) = 0.1310,$$

calculate the test statistic for testing $H_0 : \beta_1 = 0$.

T? What distribution does the test statistic follow?
Under what assumptions? Do you reject H_0?

Example: Hypothesis testing

Since $\beta_1^* = 0$, and since the data was collected over 93 years (93 samples), we calculate the observed value of the test statistic as follows (denoted with lowercase t):

$$t = \frac{0.20335 - 0}{0.1310} = 1.553.$$

Assuming that $\epsilon_i | X = x$ iid $\mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, 93$. Under the null hypothesis our test statistic T follows the t_{91} distribution.

Given that the two-sided p-value is:

$$P(|T| \geq |t| \, \big| H_0) = P(|T| \geq 1.553) = 0.124,$$

is there evidence against the null hypothesis that the early and late season snowfalls are independent?

Analysis of variance

An alternative way to address the hypothesis

$$H_0 : \beta_1 = 0 \text{ (and } H_A : \beta_1 \neq 0).$$

is via comparing the fit of two regression models.

Model	RSS
$y_i = \beta_0 + \epsilon_i$	$\sum_{i=1}^n (y_i - \hat{\beta}_0)^2 = \sum_{i=1}^n (y_i - \overline{y})^2 = SYY$
$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	$SYY - \frac{SXY^2}{SXX} = SYY - SS_{reg}$

Analysis of variance

Formally, the hypothesis test for comparing the two models is:

$$H_0 : \mathrm{E}[Y|X = x] = \beta_0,$$
$$H_A : \mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x.$$

ANOVA Table

Source	df	SS	MS	F	p-value
Regression	1	SS_{reg}	$SS_{reg}/1$	$MS_{reg}/\hat{\sigma}^2$	
Residual	n-2	RSS	$\hat{\sigma}^2 = \frac{RSS}{n-2}$		
Total	n-1	SYY			

The mean square column is obtained by dividing the sum of squares (SS) by its corresponding degrees of freedom (df).

Analysis of variance

If the errors $\epsilon_i|X = x$ iid $\mathcal{N}(0, \sigma^2)$, then the F-statistic:

$$F = \frac{SS_{reg}/1}{\hat{\sigma}^2}$$

follows the $F_{1,n-2}$ distribution, where 1 and $n-2$ are the degrees of freedom associated with the numerator and the denominator of the F-statistic (Cochran's Theorem). See also Linear Models Handout on Canvas (not on exam).

Example: Analysis of variance

Example: Ft. Collins snowfall data (n = 93). Given:

$$SXX = 10954.069,$$
$$SXY = 2229.014,$$
$$SYY = 17572.408,$$

fill in the ANOVA table and obtain the F-test for testing the hypothesis:

$$H_0 : \mathrm{E}[Y|X = x] = \beta_0,$$
$$H_A : \mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x.$$

ANOVA Table

Source	df	SS	MS	F	p-value
Regression					
Residual					
Total		17572.408			

Example: Analysis of variance

Example: Ft. Collins snowfall data (n = 93).

ANOVA Table

Source	df	SS	MS	F	p-value
Regression	1	453.5759	453.5759	2.4111	0.1239
Residual	91	17188.83	118.1190		
Total	92	17572.408			

$$SS_{reg} = \frac{SXY^2}{SXX} = \frac{2229.014^2}{10954.069} = 453.5759$$
$$RSS = SYY - SS_{reg} = 17188.83$$
$$\frac{RSS}{91} = 188.1190 = \hat{\sigma}^2$$
$$F = \frac{453.5759}{188.1190} = 2.4111, P(F^* \geq 2.4111) = 0.1239, \text{ where } F^* \sim F_{1,91}.$$

Example: Analysis of variance

What do we conclude for testing the hypothesis

$$H_0 : \mathrm{E}[Y|X=x] = \beta_0,$$
$$H_A : \mathrm{E}[Y|X=x] = \beta_0 + \beta_1 x.$$

Is there evidence against the null?

Note: the p-value for the F-statistic in this example is the same as the p-value for the t-statistic testing $H_0 : \beta_1 = 0$ ($H_A : \beta_1 \neq 0$) in the earlier example with the Ft. Collins snowfall data. Not surprising since:

$$F = \frac{SS_{reg}}{\hat{\sigma}^2} = \frac{SXY^2}{\hat{\sigma}^2 SXX} = \frac{\hat{\beta}_1^2}{SE(\hat{\beta}_1|X=x)^2} = T^2.$$

Note on reporting p-values: It is better to report a p-value and let the reader decide whether the result is significant,rather than to simply report significance at some pre-determined level.

Recall: Confidence Intervals

Because $(\hat{\beta}_0, \hat{\beta}_1)$ follow a bivariate normal distribution, when σ^2 is known, the marginal distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are univariate normal.

$$\hat{\beta}_0|X=x \sim \mathcal{N}(\beta_0, \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{SXX})),$$

given that $\epsilon_i|X=x$ iid $\mathcal{N}(0,\sigma^2)$, $i = 1, \ldots, n$. Then

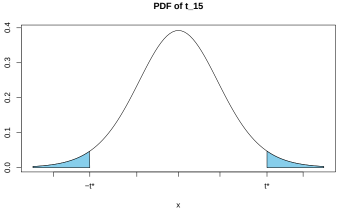
$$\frac{\hat{\beta}_0 - \beta_0}{\sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{SXX})}|X=x \sim \mathcal{N}(0,1).$$

Since σ^2 is usually not known and is instead estimated as $\hat{\sigma}^2$,

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}^2(\frac{1}{n} + \frac{\overline{x}^2}{SXX})}|X=x \sim t_{n-2}.$$

The t-distribution with $n-2$ degrees of freedom is the appropriate reference distribution for constructing the confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$.

$n = 17$ and we are interested in a 90% CI for β_0



$P(\frac{\hat{\beta}_0-\beta_0}{\sqrt{\hat{\sigma}^2/SXX}} \le |t^*| \Big| X = x) = 0.9$, so $t^* = t_{0.95,15}$. Then

$$P(-t^* \le \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0|X = x)} \le t^*) = 0.9.$$

Confidence Interval for $\hat{\beta}_0$

Since

$$P(-t^* \le \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0|X = x)} \le t^* | X = x) = 0.9,$$

$$P(\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X = x) \le \beta_0 \le \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X = x)|X = x) = 0.9,$$

a 90% confidence interval for $\hat{\beta}_0$ when $n = 17$ is:

$$\left[\hat{\beta}_0 - t^* \cdot SE(\hat{\beta}_0|X = x), \hat{\beta}_0 + t^* \cdot SE(\hat{\beta}_0|X = x)\right]$$

The general form of a two-sided $(1-\alpha) \times 100\%$ confidence interval for a symmetric probability distribution is:

Estimate ± $(1-\alpha/2)$-quantile of the prob. dist. × SE of estimate.

The interpretation of confidence intervals is based on repeated sampling. If samples of size n are drawn repeatedly and, say, 95% confidence intervals are estimated for the intercept, then 95% of those intervals (on average) would contain the true parameter β_0.

Duality: Confidence intervals and hypothesis testing

A $(1-\alpha) \times 100\%$ confidence interval for $\hat{\beta}_0$ is the set of points β_0^* such that

$$\hat{\beta}_0 - t_{1-\alpha/2,n-2} \cdot SE(\hat{\beta}_0) \leq \beta_0^* \leq \hat{\beta}_0 + t_{1-\alpha/2,n-2} \cdot SE(\hat{\beta}_0),$$

Any such β_0^* represents the null hypothesis that would not be rejected at the $100 \times \alpha\%$:

$$H_0 : \beta_0 = \beta_0^* \text{ (and } H_A : \beta_0 \neq \beta_0^*).$$

Confidence regions

So far we have only considered confidence intervals and hypothesis tests for **individual parameters**.

Often, we are interested in obtaining **simultaneous confidence intervals** for all the parameters we are estimating.

In a simple linear regression that means constructing a confidence region for $(\hat{\beta}_0, \hat{\beta}_1)$. Recall that $(\hat{\beta}_0, \hat{\beta}_1)$ follows a bivariate normal distribution, when σ^2 is known.

When σ^2 is estimated, we can construct a confidence region for $(\hat{\beta}_0, \hat{\beta}_1)$ using the Scheffé method. The reference distribution will be $F_{2,n-2}$. We will not discuss the details now.

In R, we can use functions `confint(.)` and `confidenceEllipse(.)` to obtain the confidence intervals and regions.

Example: Snow geese

Example: Consider the regression of photographic count on observer's estimate (snow geese example). Obtain 95% confidence region for the slope and intercept estimates.

```
>summary(lm(photo~obs))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1712     3.9266   0.553    0.588
obs           0.9957     0.1380   7.214 2.07e-06
---
Residual standard error: 5.804 on 16 degrees of freedom
Multiple R-squared:  0.7648,Adjusted R-squared:  0.7501
F-statistic: 52.04 on 1 and 16 DF,  p-value: 2.066e-06

> qt(0.975,16)
[1] 2.119905
```


Hypothesis testing ANOVA Confidence regions Fitted value, C.I.s and prediction C.I.s Model fit

Example: Snow geese

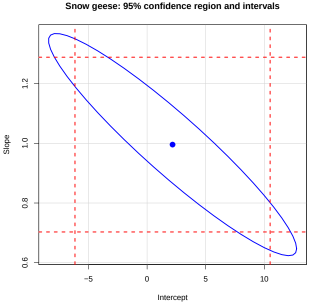
The estimates of $(\hat{\beta}_0, \hat{\beta}_1)$ and their standard errors are

$$\hat{\beta}_0 = 2.1712, SE(\hat{\beta}_0|X = x) = 3.9266, t_{0.975,16} = 2.12$$
$$\hat{\beta}_1 = 0.9957, SE(\hat{\beta}_1|X = x) = 0.1380, t_{0.975,16} = 2.12$$

Let's construct:

- the 95% confidence interval for β_0,
- the 95% confidence interval for β_1,
- the 95% joint **confidence region** for (β_0, β_1).

Example: Snow geese



Example: Snow geese

The confidence region has the shape of an ellipse.

The dashed lines show confidence intervals for each parameter.

Notice that these lines do not enclose the ellipse exactly (if they did, they would be jointly correct confidence intervals).

We are interested to test the null hypothesis which says the observer's count is perfect (the same as the photographic count).

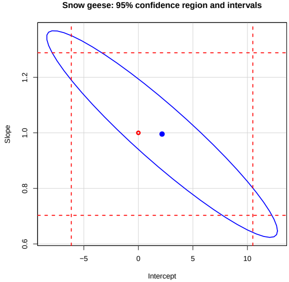
Can you write down this null hypothesis formally?

The null hypothesis of perfect observers count:

$$H_0 : (\beta_0, \beta_1) = (0, 1) \text{ versus } H_A : (\beta_0, \beta_1) \neq (0, 1)$$

Let's plot the value of the null.

Example: Snow geese



Can we reject the null hypothesis?

Example: Snow geese

The point value for our hypothesis:

$$(\beta_0, \beta_1) = (0, 1)$$

lies within the ellipse and within the 95% confidence intervals for the intercept and slope.

Hence, we cannot reject the null hypothesis that observer's count is exact.

It is possible for the point of interest to lie outside of the ellipse, but within the individual confidence interval.
Would you reject the above H_0 in that case?

It is also possible for the point of interest to lie within the ellipse but outside of the confidence intervals.
Would you reject the above H_0 in that case?

Fitted values

Given a new predictor value, x_* what is the fitted value?

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

This fitted value is the predicted mean of the response y at the value x^*. We need to distinguish between predictions of the mean of the response (**fitted values**, see above) and the value of the response.

Note that are model assumes:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where ϵ_i iid $\mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$. Hence, for a value x_*, $y_* = \beta_0 + \beta_1 x_* + \epsilon_*$, where we do not observe ϵ_*.

This difference affects how we construct *fitted value confidence intervals* and *prediction confidence intervals*. (prediction intervals).

Variance and Bias of fitted values in OLS

The **fitted values** are values on the regression line

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*.$$

The uncertainty in the fitted value comes from the uncertainty in the estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$. Based on the previous lecture, we know that for least squares estimates:

$$E[\hat{y}_* | X = x_*] = \beta_0 + \beta_1 x_*$$

The least squares fitted value is an **unbiased** estimate of the mean.
The **variance** for the fitted least squares estimate is:

$$\begin{aligned}
&\text{Var}[\hat{y}_* | X = x_*] = \text{Var}[\hat{\beta}_0 + \hat{\beta}_1 x_* | X = x_*] \\
&= \text{Var}[\hat{\beta}_0 | X = x_*] + x_*^2 \text{Var}[\hat{\beta}_1 | X = x_*] + 2 x_* \text{Cov}[\hat{\beta}_0, \hat{\beta}_1 | X = x_*] \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right) + \sigma^2 x_*^2 \frac{1}{SXX} - 2\sigma^2 x_* \frac{\bar{x}}{SXX} = \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX} \right).
\end{aligned}$$

Fitted values

Then

$$SE(\hat{y}_*|X = x_*) = \hat{\sigma}\Big(\frac{1}{n} + \frac{(x_* - \overline{x})^2}{SXX}\Big)^{1/2}$$

The confidence interval for the fitted value of y given x can be constructed as

$$\hat{y}_*|_{x_*} \pm t_{1-\alpha/2,n-2} \cdot SE(\hat{y}_*|X = x_*).$$

Note that the confidence interval for the fitted value is wider the further away we are from $\overline{x}$.

Predicted values

Since the true value of y_* according to our model is

$$y_* = \beta_0 + \beta_1 x_* + \epsilon_*,$$

As before:

$$\mathrm{E}[y_* - \hat{y}_*|X = x_*] = 0.$$

What about $\mathrm{Var}[y_* - \hat{y}_*|X = x_*]$? How far away is our predicted (fitted) value from the actual value y_*? Using the formula for the variance of the sum of two uncorrelated variables, we obtain:

$$\mathrm{Var}[y_* - \hat{y}_*|X = x_*] = \mathrm{Var}[\beta_0 + \beta_1 x_* + \epsilon_* - \hat{y}_*|X = x_*]$$

$$= \mathrm{Var}[\epsilon_*|X = x_*] + \mathrm{Var}[\hat{y}_*|X = x_*] = \sigma^2 + \sigma^2\Big(\frac{1}{n} + \frac{(x_* - \overline{x})^2}{SXX}\Big)$$

$$= \sigma^2\Big(1 + \frac{1}{n} + \frac{(x_* - \overline{x})^2}{SXX}\Big).$$

Compare: Uncertainty in fitted and predicted values

Since $\mathrm{Var}[y_* - \hat{y}_* | X = x_*] = \sigma^2(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})$ The standard error is:

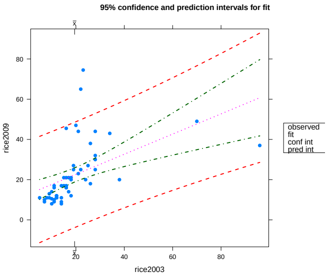
$$SE(y_* - \hat{y_*} | X = x_*) = \hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}\right)^{1/2}$$

The prediction interval (predicted observation value confidence interval) for y_* can be constructed as

$$\hat{y}_*|_{x_*} \pm t_{1-\alpha/2, n-2} \cdot SE(y_* - \hat{y_*} | X = x_*).$$

The prediction interval for y_* is always wider than the confidence interval for $\hat{y}_*$.

Compare: Uncertainty in fitted and predicted values



Example: snowgeese data.

R^2: the Coefficient of Determination

R^2 is the proportion of variability in the response that is explained by the regression.

$$R^2 = \frac{SS_{reg}}{SYY} = 1 - \frac{RSS}{SYY}.$$

It takes values in $[0,1]$ and is a scale-free summary of the strength of linear relationship between the x's and the y's in the data.
Since $SS_{reg} = \frac{SXY^2}{SXX}$, we can also write

$$R^2 = \frac{SXY^2}{SXX \cdot SYY} = r_{XY}^2.$$

Hence, R^2 can be thought of as the square of the sampling correlation between the predictor and the response.
R^2 is a measure of goodness of fit of a linear regression.

Example: Snow geese

Calculate R^2 from:

```
> anova(lm(photo~obs))
Analysis of Variance Table

Response: photo
          Df  Sum Sq Mean Sq F value    Pr(>F)
obs        1 1752.70 1752.70  52.037 2.066e-06
Residuals 16  538.91   33.68
---
> 1752.70/(1752.70+538.91)
[1] 0.7648335
```

Model Fit

An alternative measure of (absolute) fit is $\hat{\sigma}$.

While R^2 is scale-free, $\hat{\sigma}$ is measured in the units of the response.

This can be both an advantage and a disadvantage: one must understand the practical significance of $\hat{\sigma}$ in order to interpret its value.

Mean Squared Error

Another way to asses model fit is using the **generalization error** (**mean squared error of the estimator**)
$MSE(\hat{y}) = \mathrm{E}[(y-\hat{y})^2|X=x]$.

It can be shown that:

$$MSE(\hat{y}) = (\mathrm{E}[\hat{y}|X=x]-y)^2 + \mathrm{Var}[\hat{y}|X=x] = Bias(\hat{y})^2 + \mathrm{Var}[\hat{y}|X=x].$$

This is known as the bias-variance trade-off.
For least squares estimates:

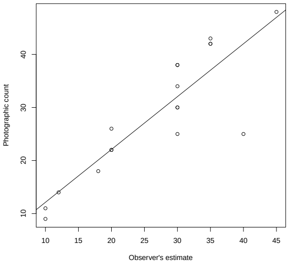
$$MSE(\hat{y}) = \sigma^2\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{SXX}\right).$$

The **estimated (within sample) mean squared error** computed as

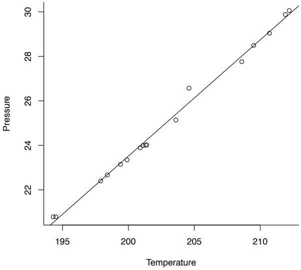
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

We will revisit the MSE later in the course.

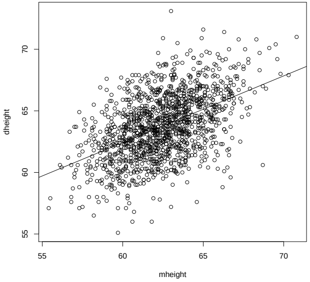
Snow geese: R-squared=0.7648, $\hat{\sigma}$ = 5.804, MSE=29.94



Forbes data: R-squared=0.9944, $\hat{\sigma}$ = 0.2328, MSE = 0.048



Heights: R-squared=0.2408, $\hat{\sigma}$ = 2.266, MSE = 5.129

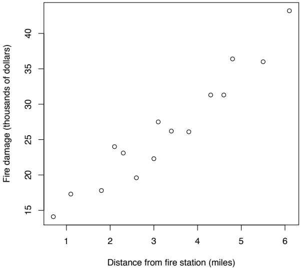


Example: Interpretation of the slope

Example: Fire damage.
Consider a large suburb of a major city.
Is the amount of fire damage related to the proximity of the nearest fire station?
Let y be the amount of fire damage in thousands of dollars and x be the distance to the nearest fire station in miles.
A sample of 15 recent residential fires was selected.

Data: `fire.df` in R package `s20x`.

Example: Fire damage



Example: Fire damage

Fitting the regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ with } \epsilon_i \sim \mathcal{N}(0, \sigma^2), \epsilon_i \text{ iid.}$$

we obtain the following (partial) R output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2779     1.4203   7.237 6.59e-06
distance      4.9193     0.3927  12.525 1.25e-08
---
Residual standard error: 2.316 on 13 degrees of freedom
Multiple R-squared:  0.9235,Adjusted R-squared:  0.9176
F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

Example: Interpretation of the slope

We can provide the usual interpretation for $\hat{\beta}_1$:

for every additional mile of distance to the nearest fire station, damage from residential fires increases by 4.9 thousand of dollars on average.

True or false?

- If we take a number of houses at random and move them an additional mile away from the nearest fire station, we would expect their fire damage to increase by 4.9 thousand of dollars on average.

False!
Note: Observational studies cannot be used to infer causal relationship without additional information external to the study.