

Mathematics for Machine Learning

MAD-B3-2526-S2-MAT0611

Hypothesis Testing in Regression Models

Agenda

1. Hypothesis Testing Recap
2. Hypothesis Testing Foundations
3. Model Fitting with Statsmodels
4. Testing Individual Coefficients (t-tests)
5. Testing Overall Model Significance (F-tests)
6. Multiple Testing Considerations

Hypothesis Testing Recap Example

Testing Population Mean Height

Research Question: Is the average height of adults in a population equal to 170 cm?

Setup:

- Collect sample of $n = 50$ adults
- Measure heights: $\bar{x} = 172.5$ cm, $s = 8.2$ cm
- Population standard deviation unknown

Hypotheses (two-sided test):

$$H_0 : \mu = 170 \quad \text{vs.} \quad H_1 : \mu \neq 170$$

Distribution of the Sample Mean

Sampling Distribution

When we take repeated samples from a population and compute \bar{X} for each sample, the distribution of \bar{X} is called the **sampling distribution of the sample mean**.

Key Properties:

If X_1, X_2, \dots, X_n are i.i.d. with mean μ and variance σ^2 :

$$E[\bar{X}] = \mu$$
$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \Rightarrow \text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Note that both the variance and standard error decrease as n increases.

Exercise: Prove these properties.

Central Limit Theorem (CLT)

For large enough sample size n , the sampling distribution of \bar{X} converges in distribution to a normal distribution:

$$\bar{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Equivalently, the standardized sample mean converges to a standard normal:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

Key insight: This result holds regardless of the original population distribution of X_i (provided the variance is finite).

When Population Variance is Unknown

Problem: In practice, we rarely know σ^2 .

Solution: Replace σ with the sample standard deviation s . This introduces additional uncertainty, so the test statistic follows a t-distribution:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

t-distribution Properties:

- Similar to standard normal but with heavier tails
- Converges to $N(0, 1)$ as $n \rightarrow \infty$
- Degrees of freedom: $n - 1$
- Valid when X_1, X_2, \dots, X_n are normally distributed OR n is large enough for CLT (typically $n \geq 30$)

t-test Computation

Looking back at our hypothesis test example, we can calculate the **test statistic**:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{172.5 - 170}{8.2/\sqrt{50}} = \frac{2.5}{1.16} = 2.16$$

Degrees of freedom: $df = n - 1 = 49$

Critical values (at $\alpha = 0.05$):

- $t_{0.025,49} = \pm 2.01$

Decision:

- Since $|t| = 2.16 > 2.01$, reject H_0
- **p-value** = $2 \cdot P(T > 2.16) \approx 0.035 < 0.05$

Conclusion: There is evidence that the population mean height differs from 170 cm.

Python Implementation

```
import numpy as np
from scipy import stats

# Sample data
np.random.seed(42)
heights = np.random.normal(172.5, 8.2, 50) # this would be your sample data

# Compute statistics
x_bar = np.mean(heights)
s = np.std(heights, ddof=1)
n = len(heights)
mu_0 = 170

# t-test
t_stat = (x_bar - mu_0) / (s / np.sqrt(n))
p_value = 2 * (1 - stats.t.cdf(abs(t_stat), df=n-1))

print(f"Sample mean: {x_bar:.2f} cm")
print(f"Sample std: {s:.2f} cm")
print(f"t-statistic: {t_stat:.3f}")
print(f"p-value: {p_value:.4f}")

# Using scipy's ttest_1samp
t_stat_scipy, p_value_scipy = stats.ttest_1samp(heights, mu_0)
print(f"\nUsing scipy.stats.ttest_1samp:")
print(f"t-statistic: {t_stat_scipy:.3f}")
print(f"p-value: {p_value_scipy:.4f}")
```


Hypothesis Testing Foundations

The Basic Framework

Null Hypothesis (H_0): The claim we test (typically "no effect" or "no difference")

Alternative Hypothesis (H_1 or H_a): What we conclude if H_0 is rejected

Test Statistic: A function of the data that measures evidence against H_0

p-value: The probability of observing a test statistic at least as extreme as the one observed, assuming H_0 is true

Significance Level (α): Threshold for rejection (commonly 0.05)

Decision Rule

Reject H_0 if:

- p-value $< \alpha$, or equivalently
- The test statistic falls in the rejection region

Fail to reject H_0 if:

- p-value $\geq \alpha$

Important Notes:

- "Fail to reject" H_0 does not mean we "accept" H_0
- Absence of evidence is not evidence of absence

Type I and Type II Errors

	H_0 True	H_0 False
Reject H_0	Type I Error (False Positive)	Correct (True Positive)
Don't Reject H_0	Correct (True Negative)	Type II Error (False Negative)

- **Type I Error:** $P(\text{Reject } H_0 | H_0 \text{ true}) = \alpha$
- **Type II Error:** $P(\text{Don't Reject } H_0 | H_0 \text{ false}) = \beta$
- **Power:** $1 - \beta$ (probability of correctly rejecting false H_0)

One-sided vs. Two-sided Tests

Two-sided (most common):

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0$$

$$\text{p-value} = 2 \cdot P(T > |t|)$$

One-sided (directional):

$$H_0 : \mu \leq 0 \quad \text{vs.} \quad H_1 : \mu > 0$$

$$\text{p-value} = P(T > t)$$

When to use one-sided tests:

- Strong prior belief about the direction of the effect
- Only one direction is practically or theoretically meaningful

Confidence Intervals

A $(1 - \alpha) \times 100\%$ confidence interval for μ :

$$\hat{\mu} \pm t_{\alpha/2, n-1} \cdot \text{SE}(\hat{\mu})$$

Correct Interpretation:

- If we repeated the sampling procedure many times, $(1 - \alpha) \times 100\%$ of the computed intervals would contain the true value of μ
- **Incorrect interpretation:** "There is a 95% probability that μ is in this specific interval"

Connection to Hypothesis Testing:

- If the confidence interval does not contain 0, then we reject $H_0 : \mu = 0$ at significance level α

Fitting Models with Statsmodels

```
import numpy as np
import pandas as pd
import statsmodels.api as sm

# Generate sample data
np.random.seed(42)
n = 100
X1 = np.random.randn(n)
X2 = np.random.randn(n)
Y = 2 + 3*X1 - 1.5*X2 + np.random.randn(n)

# Prepare design matrix (add constant)
X = pd.DataFrame({'X1': X1, 'X2': X2})
X = sm.add_constant(X)

# Fit model
model = sm.OLS(Y, X).fit()
print(model.summary())
```

OLS Regression Results

```

=====
Dep. Variable:          y      R-squared:          0.913
Model:                  OLS    Adj. R-squared:       0.911
Method:                 Least Squares    F-statistic:       506.0
Date:                   Wed, 21 Jan 2026    Prob (F-statistic): 4.79e-52
Time:                   10:00:00    Log-Likelihood:    -147.62
No. Observations:      100    AIC:              301.2
Df Residuals:          97    BIC:              309.1
Df Model:               2
Covariance Type:       nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	2.0886	0.108	19.298	0.000	1.874	2.303
X1	3.2261	0.120	26.859	0.000	2.988	3.464
X2	-1.5123	0.114	-13.221	0.000	-1.739	-1.285

```

=====
Omnibus:                3.125    Durbin-Watson:          2.220
Prob(Omnibus):          0.210    Jarque-Bera (JB):       3.080
Skew:                   0.108    Prob(JB):               0.214
Kurtosis:               3.832    Cond. No.                1.23
=====

```

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

What Do These Numbers Mean?

For each coefficient:

- `coef` : estimated value $\hat{\beta}_j$
- `std err` : standard error $SE(\hat{\beta}_j)$
- `t` : t-statistic
- `P>|t|` : p-value for testing $H_0 : \beta_j = 0$
- `[0.025, 0.975]` : 95% confidence interval

Overall model:

- `R-squared` : proportion of variance explained
- `F-statistic` : test of overall model significance
- `Prob (F-statistic)` : p-value for F-test
- `AIC` , `BIC` : information criteria for model selection

Testing Individual Coefficients: t-tests

Hypothesis

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

Interpretation: Does predictor X_j have a significant effect on the response?

Test Statistic

$$t = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$$

Under H_0 : $t \sim t_{n-p-1}$ (t-distribution with $n - p - 1$ degrees of freedom)

Standard Error of Coefficients

For **linear regression**, the Fisher Information Matrix is:

$$\mathbf{I}(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X}$$

The standard error of $\hat{\beta}_j$ is:

$$\text{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 [(\mathbf{X}^T \mathbf{X})^{-1}]_{jj}}$$

where $\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1}$ is the estimated residual variance.

Key Factors Affecting Standard Error:

- Residual variance σ^2 : larger variance \Rightarrow larger SE
- Sample size n : larger sample \Rightarrow smaller SE
- Multicollinearity: higher correlation among predictors \Rightarrow larger SE

t-test: Computation

Example from our model output:

For X_1 :

- $\hat{\beta}_1 = 3.2261$
- $\text{SE}(\hat{\beta}_1) = 0.120$
- $t = \frac{3.2261}{0.120} = 26.859$

p-value: $P(|T| > 26.859)$ where $T \sim t_{97}$

Since $t = 26.859$ is extremely large:

- p-value ≈ 0.000 (very strong evidence against H_0)
- **Conclusion:** Reject H_0 ; X_1 is highly significant

Testing Overall Model Significance: F-test

Hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \neq 0$$

Interpretation: Does the model explain variation in the response beyond what the intercept alone explains?

F-statistic

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)} = \frac{\text{MSR}}{\text{MSE}}$$

where:

- TSS = Total Sum of Squares = $\sum_{i=1}^n (y_i - \bar{y})^2$
- RSS = Residual Sum of Squares = $\sum_{i=1}^n (y_i - \hat{y}_i)^2$
- MSR = Mean Square Regression = $(\text{TSS} - \text{RSS})/p$
- MSE = Mean Square Error = $\text{RSS}/(n - p - 1)$

Under H_0 : $F \sim F_{p, n-p-1}$

ANOVA Table

Source	SS	df	MS	F
Regression	TSS - RSS	p	MSR	MSR/MSE
Residual	RSS	$n - p - 1$	MSE	
Total	TSS	$n - 1$		

Decomposition:

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{Total Variation}} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{Explained}} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{Unexplained}}$$

R-squared and Adjusted R-squared

R-squared (Coefficient of Determination):

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

- Represents the proportion of variance in the response explained by the model
- Range: $[0, 1]$ (higher values indicate better fit)
- **Limitation:** Always increases (or stays constant) when adding predictors, even if irrelevant

Adjusted R-squared:

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n - p - 1)}{\text{TSS}/(n - 1)}$$

- Penalizes model complexity

Relationship: F-test and R-squared

$$F = \frac{R^2 / p}{(1 - R^2) / (n - p - 1)}$$

Key Insights:

- Large R^2 implies large F , which leads to rejecting H_0
- The F-test accounts for both sample size and the number of predictors
- A significant F-test does not imply that all individual coefficients are significant

Example: F-test Interpretation

From our model output:

```
F-statistic:          506.0
Prob (F-statistic):    4.79e-52
```

Interpretation:

- $F = 506.0$ is very large
- $p\text{-value} = 4.79 \times 10^{-52} < 0.05$
- **Conclusion:** There is very strong evidence that at least one predictor is significant
- The model explains substantial variation in Y beyond the intercept-only model

Model Selection Criteria

Akaike Information Criterion (AIC):

$$\text{AIC} = -2\ell(\hat{\beta}) + 2k$$

Bayesian Information Criterion (BIC):

$$\text{BIC} = -2\ell(\hat{\beta}) + k \log(n)$$

where k is the number of parameters and n is the sample size.

Usage Guidelines:

- Lower values indicate better models
- BIC penalizes model complexity more heavily than AIC (especially for large n)
- Both can be used for comparing non-nested models
- AIC aims to minimize prediction error; BIC approximates Bayes factors

Logistic Regression: Wald Test

For **logistic regression**, we use the **Wald test**:

$$z = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$$

Under $H_0 : \beta_j = 0$: $z \stackrel{\text{approx}}{\sim} N(0, 1)$ (for large n)

For **logistic regression**, the Fisher Information Matrix is:

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

where $\mathbf{W} = \text{diag}(\hat{p}_i(1 - \hat{p}_i))$ and $\hat{p}_i = \frac{1}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}}$

The standard error of $\hat{\beta}_j$ is:

$$\text{SE}(\hat{\beta}_j) = \sqrt{[\mathbf{I}(\hat{\boldsymbol{\beta}})^{-1}]_{jj}} = \sqrt{[(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}]_{jj}}$$

Key Differences from Linear Regression:

- No separate σ^2 parameter (variance is determined by the Bernoulli distribution)
- Weights \mathbf{W} depend on predicted probabilities \hat{p}_i , not on observed responses y_i
- The Hessian is $\mathbf{H} = -\mathbf{X}^T \mathbf{W} \mathbf{X}$, so $\mathbf{I}(\boldsymbol{\beta}) = -\mathbf{H}$
- Wald test uses the standard normal distribution (asymptotically), not the t-distribution, because the variance structure is fully specified by the model

Logistic Regression Example

```
# Generate binary outcome
np.random.seed(42)
n = 200
X1 = np.random.randn(n)
X2 = np.random.randn(n)
z = -0.5 + 2*X1 + 1.5*X2
p = 1 / (1 + np.exp(-z))
Y_binary = np.random.binomial(1, p)

# Fit logistic regression
X = pd.DataFrame({'X1': X1, 'X2': X2})
X = sm.add_constant(X)
logit_model = sm.Logit(Y_binary, X).fit()
print(logit_model.summary())
```

Logistic Regression Output

Logit Regression Results						
=====						
Dep. Variable:	y	No. Observations:	200			
Model:	Logit	Df Residuals:	197			
Method:	MLE	Df Model:	2			
Date:	Mon, 21 Jan 2026	Pseudo R-squ.:	0.4521			
Time:	10:00:00	Log-Likelihood:	-67.234			
converged:	True	LL-Null:	-122.73			
Covariance Type:	nonrobust	LLR p-value:	3.456e-25			
=====						
	coef	std err	z	P> z	[0.025	0.975]

const	-0.5423	0.178	-3.047	0.002	-0.891	-0.194
X1	2.1234	0.289	7.346	0.000	1.557	2.690
X2	1.6789	0.256	6.558	0.000	1.178	2.180
=====						

Interpreting Logistic Regression Coefficients

Coefficient interpretation:

$$\exp(\beta_j) = \text{odds ratio}$$

- One unit increase in X_j multiplies odds by $\exp(\beta_j)$

Example: $\hat{\beta}_1 = 2.1234$

- Odds ratio = $\exp(2.1234) = 8.36$
- One unit increase in X_1 multiplies odds of $Y = 1$ by 8.36

Significance:

- $z = 7.346$, p-value < 0.001
- Strong evidence that X_1 affects outcome

Resources

- **Books:**

- Wasserman: *All of Statistics* (Ch. 10-11)
- James et al.: *Introduction to Statistical Learning* (Ch. 3)
- Greene: *Econometric Analysis* (Ch. 4-5)

- **Python Libraries:**

- statsmodels: comprehensive statistical models
- scikit-learn: machine learning implementations
- scipy.stats: statistical tests