Simple Linear Regression

Part 3: Inference

STAT 705: Regression and Analysis of Variance

Variability of the Estimators

- Estimators: $\hat{\beta}_0$ and $\hat{\beta}_1$
- Calculated from the observed (X, Y) pairs in the sample
- This implies that each estimator
  - is a random variable
  - has a probability distribution
  - has a mean and a standard deviation

Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

	Mean	**Variance**
Intercept	$E(\hat{\beta}_0) = \beta_0$	$Var\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_{XX}}\right)$
Slope	$E(\hat{\beta}_1) = \beta_1$	$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{SS_{XX}}$

For both $\hat{\beta}_0$ and $\hat{\beta}_1$, the standard error is the square root of variance.

Since σ^2 is not known, we substitute $\hat{\sigma}^2$.

Then $\hat{\beta}_0$ and $\hat{\beta}_1$ follow a t distribution with $n-2$ degrees of freedom.

Standard Errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

Intercept	Slope
Point estimate $= \hat{\beta}_0 = -21$	Point estimate $= \hat{\beta}_1 = 35.7$
$\hat{Var}\left(\hat{\beta}_0\right) = MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_{XX}}\right)$	$\hat{Var}\left(\hat{\beta}_1\right) = \frac{MSE}{SS_{XX}}$
$\hat{Var}\left(\hat{\beta}_0\right) = 7867.2\left(\frac{1}{11} + \frac{18.5^2}{643.56}\right)$	$\hat{Var}\left(\hat{\beta}_1\right) = \frac{7867.2}{643.56} = 12.22$
$\hat{Var}\left(\hat{\beta}_0\right) = 4899.69$	$se\left(\hat{\beta}_1\right) = \sqrt{12.22} = 3.5$
$se\left(\hat{\beta}_0\right) = \sqrt{4899.69} = 70.0$	

Confidence Intervals for β_0 and β_1

- General form of confidence interval is

  (point estimate) $\pm$ (critical value)*(standard error)
- The critical value is from the t distribution with $n - 2$ degrees of freedom (n is the number of <u>pairs</u> in the data)
- For Lead vs. Traffic example, with α = .05
  $df = n - 2 = 9$ ⇨ critical value = $t_{\alpha/2,\,9} = 2.262$
- 95% CI for β_0 is
  $-21 \pm (2.262)*70.0$, or (-179.3, 137.3)
- 95% CI for β_1 is
  $35.7 \pm (2.262)*3.5$, or (27.8, 43.6)

Hypothesis Test for β_1

For some constant c, test H$_0$: $\beta_1 = c$ vs. H$_a$: $\beta_1 \neq c$

$$\text{Test statistic} = \frac{\hat{\beta}_1 - c}{se\left(\hat{\beta}_1\right)} \qquad \text{Critical value} = t_{\alpha/2,\, df=n-2}$$

Reject H_0 if $\left|\text{Test statistic}\right| > \text{Critical value}$

Example: Lead vs. Traffic, test H$_0$: $\beta_1 = 0$ vs. H$_a$: $\beta_1 \neq 0$

$$\text{Test statistic} = \frac{35.7 - 0}{3.5} = 10.2 \qquad \text{Critical value} = 2.262$$

$\left|10.2\right| > 2.262$, so we reject H_0

The sample provides evidence that the slope is not 0.

Interpret the Inference on Slope

- In the previous hypothesis test, we concluded that the slope of the line regressing Lead on Traffic is not 0. This means
  - the predictor (Traffic) DOES help explain response (Lead)
  - the predictor should be kept in the model
- If the conclusion had been opposite (i.e., do not reject H_0), then any of the following could be plausible
  - the ‘true’ slope is 0
  - the ‘true’ regression model does not include the predictor
  - a reduced model ($Y = \beta_0 + \varepsilon$) may be adequate
  - the ‘true’ regression model is not linear

Hypothesis Test for β_0

For some constant c, test H$_0$: $\beta_0 = c$ vs. H$_a$: $\beta_0 \neq c$

$$\text{Test statistic} = \frac{\hat{\beta}_0 - c}{se\left(\hat{\beta}_0\right)} \qquad \text{Critical value} = t_{\alpha/2,\, df=n-2}$$

$$\text{Reject H}_0 \text{ if } \left|\text{Test statistic}\right| > \text{Critical value}$$

Example: Lead vs. Traffic, test H$_0$: $\beta_0 = 0$ vs. H$_a$: $\beta_0 \neq 0$

$$\text{Test statistic} = \frac{-21 - 0}{70} = -0.3 \qquad \text{Critical value} = 2.262$$

$$\left|-0.3\right| \text{ is not} > 2.262\text{, so we do not reject H}_0$$

Conclusion: It is reasonable to believe that the intercept could be 0.

Relation Between Tests and CI

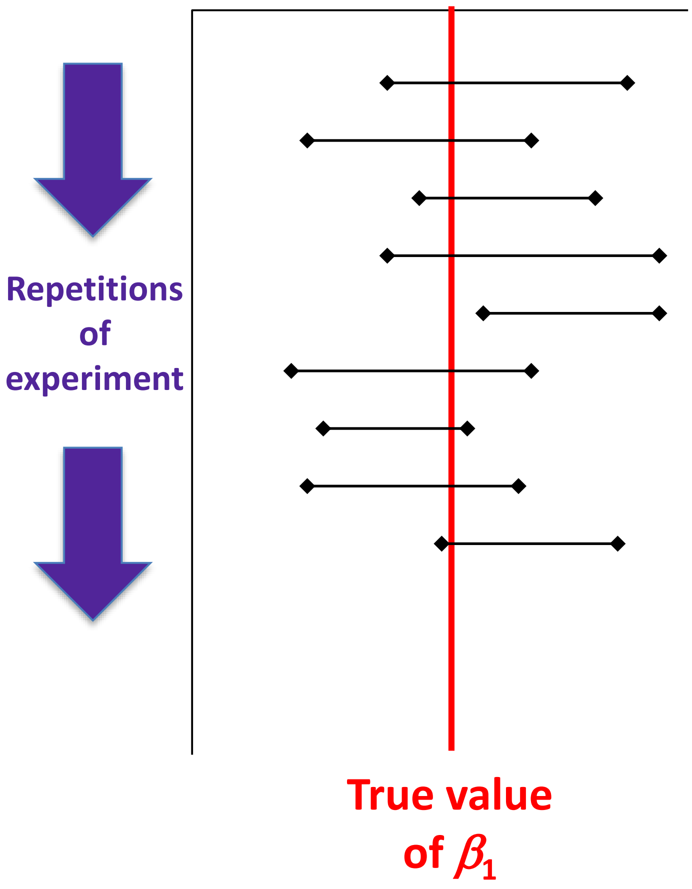
- Two-sided hypothesis tests are equivalent to confidence intervals
- For the slope:
  - At $\alpha = 0.05$, we rejected H_0: $\beta_1 = 0$ in favor of H_a: $\beta_1 \neq 0$
  - A 95% confidence interval for β_1 is (27.8, 43.6)
- The confidence interval does not contain the hypothesized value (0), so 0 is not a 'plausible' value for β_1 and we should reject the null hypothesis.

Interpreting a Confidence Interval

- A confidence interval for a population parameters (e.g., either β_0 or β_1) is based on the sampling distribution of its estimator $\left(\text{e.g., either } \hat{\beta}_0 \text{ or } \hat{\beta}_1\right)$
- We found a 95% confidence interval for β_1 is (27.8, 43.6)
- INCORRECT INTERPRETATION:
  - The probability is 95% that β_1 is between 27.8 and 43.6.
- Why is this incorrect?
  - β_1 is a population parameter. It is a fixed, but unknown, value.
  - β_1 is NOT a random variable; it does NOT have a probability distribution

Correct Interpretation a CI

- We found a 95% confidence interval for β_1 is (27.8, 43.6)
- What this means:
  - If we repeat the experiment many times
  - Generate a new confidence interval for each repetition
  - In approximately 95% of the repetitions, the confidence interval would contain the true value of the parameter
- Brief interpretation
  - The interval [27.8,43.6] contains the true value of the parameter β_1 with 95% confidence



Coefficient of Determination

- Also known as 'R-square'
- Definition: $R^2 = 1 - \dfrac{SSE}{SS_{YY}}$
- For the Lead vs. Traffic example,

$$R^2 = 1 - \frac{70{,}805.1}{892{,}522.67} = 0.921$$

- Interpretation: Approximately 92.1% of the variability in lead concentration can be explained by the traffic volume.

Prediction and Estimation

- We can use the linear regression equation to
  1. Estimate the mean value of Y for a specific X
  2. Predict the value for an individual Y with a specified X
- How are these different?

Consider all the locations in the population that have traffic volume 22 thousand vehicles.

1. We can estimate the mean lead concentration across all these locations.
2. We can predict the lead concentration for one of these locations.

Prediction and Estimation

- Estimate $E(Y \mid X_0)$, the mean Y for a specific X_0
  - point estimate = $\hat{\beta}_0 + \hat{\beta}_1 X_0$
  - variance of this estimate = $\sigma^2 \left(\frac{1}{n} + \frac{\left(X_0 - \bar{X} \right)^2}{SS_{XX}} \right)$
- Predict an individual Y for a specific X_0
  - point estimate = $\hat{\beta}_0 + \hat{\beta}_1 X_0$
  - variance of this estimate = $\sigma^2 \left(1 + \frac{1}{n} + \frac{\left(X_0 - \bar{X} \right)^2}{SS_{XX}} \right)$
- For both variances, substitute MSE for σ^2
- How are these different? How are they the same?

Example of Prediction and Estimation

- In the Lead vs. Traffic example, consider sites that have 22,000 vehicles (X_0 = 22)
- Estimated Y

  -21 + 35.7×22 = 764.4 micrograms of lead per gram of bark
- variance for the mean estimate

$$\text{MSE}\left(\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{SS_{XX}}\right)=7867.2\left(\frac{1}{11}+\frac{\left(22-18.5\right)^2}{643.56}\right)=864.95$$

- variance for the individual predicted value

$$\text{MSE}\left(1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{SS_{XX}}\right)=7867.2\left(1+\frac{1}{11}+\frac{\left(22-18.5\right)^2}{643.56}\right)=8732.15$$

Confidence and Prediction Intervals

- $X_0 = 22$, point estimate $= 764.4$, critical value $= 2.262$
- Confidence interval for the mean Lead across all locations with 22,000 vehicles

  estimate $\pm$ (critical value)$\times$se(est) $\Rightarrow$ $764.4 \pm (2.262)\times(864.95)^{1/2}$

  (697.9, 830.9)

- Prediction interval for the Lead concentration at one location that has 22,000 vehicle

  estimate $\pm$ (critical value)$\times$se(est) $\Rightarrow$ $764.4 \pm (2.262)\times(8732.15)^{1/2}$

  (553.0, 975.8)

- Prediction intervals are wider because there is more variability in a single location than in the mean across numerous locations.

Confidence and Prediction Intervals

- Now obtain the CI and PI for sites with 18,500 vehicles
- For both intervals
  - X_0 = 18.5, which happens to be the mean of X
  - Point estimate = -21 + 35.7$\times$18.5 = 639.45
  - Critical value = 2.262
- For the mean estimate
  - Variance = MSE*(1/n) = 7867.2/11 = 715.2; std.err = 26.74
  - Interval is 639.45 $\pm$ 2.262*26.74, or (579, 700)
- For the individual predicted value
  - Variance = MSE*(1 + 1/n) = 7867.2 (1 + 1/11) = 8582.4; std.err = 92.64
  - Interval: 639.45 $\pm$ 2.262*92.64, or (430, 849).

Summary: CI and PI

	Confidence Interval	Prediction Interval
Traffic = 22,000 (X_0 = 22)	698 to 831 (width = 133)	553 to 976 (width = 423)
Traffic = 18,500 (X_0 = 18.5)	579 to700 (width = 121)	430 to 849 (width = 419)

- In general,
  - The width of any of these intervals depends on the value of X_0
  - Both types of intervals are narrower when X_0 is closer to the mean than when X_0 is farther away
  - For any given X_0, confidence intervals are narrower than prediction intervals

Comparison

Confidence Interval on Mean Response	Prediction Interval for New Observation
• Mean of the distribution of Y for a given X • More precise ⇨ narrower • Inference on a parameter • Apply to a mean response	• Individual outcome ⇨ more uncertainty • Less precise ⇨ broader • Statement on a random variable • Apply to a single new observation

Things You Should Know

- Interpret confidence intervals
  - for the slope β_1
  - for the intercept β_0
  - for the mean (expected value) of the response
- Interpret prediction intervals
  - for an individual Y, given a specific X
- Interpret
  - Point estimates for σ^2, β_0 and β_1
- Difference between an estimate and a parameter
- Difference between a confidence interval and a prediction interval