Chapter 15

Simple Linear Regression

Detailed Solutions

15.1 Basic Concept

15.1 The Probabilistic Model

Problem: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
Solution:

(a) **Deterministic vs. Stochastic:**

- **Signal (Deterministic):** $\beta_0 + \beta_1 x_i$. This is the expected value $E[Y|x]$, representing the underlying trend.
- **Noise (Stochastic):** ϵ_i. This represents random variation, measurement errors, or unobserved factors.

(b) **Assumptions on ϵ (LINE):**

(a) **L**inearity: The relationship between X and Y is linear.
(b) **I**ndependence: Errors ϵ_i and ϵ_j are independent.
(c) **N**ormality: Errors are normally distributed $\epsilon \sim N(0, \sigma^2)$.
(d) **E**qual Variance (Homoscedasticity): The variance σ^2 is constant for all x.

(c) **Residuals:**

- ϵ_i: The **true** unobservable error $(Y_i - (\beta_0 + \beta_1 x_i))$.
- e_i: The **observed** residual $(Y_i - \hat{y}_i)$ calculated from the sample data. We use e_i to estimate properties of ϵ_i.

(a) Signal: $\beta_0 + \beta_1 x$, Noise: ϵ
(b) Linearity, Independence, Normality, Equal Variance
(c) ϵ is theoretical, e is calculated.

15.2 Correlation vs. Regression

Problem: $r = 0.9$ vs β_1, Extrapolation.
Solution:

(a) **Slope vs Correlation:** No, β_1 is not necessarily 0.9.

$$\beta_1 = r\left(\frac{S_y}{S_x}\right)$$

The slope depends on the units/scale of X and Y. r is unitless. $r = 0.9$ means strong positive linearity, but the slope could be 0.001 or 1000 depending on the ratio of standard deviations.

(b) **Extrapolation:** The relationship observed in the range $[10, 20]$ may not hold at $x = 100$. The process might saturate, curve, or break down (e.g., Hooke's law fails after yield point). Predictions far outside the data range are unreliable.

(a) $\beta_1 = r(S_y/S_x)$. Not equal.
(b) Relationship may change outside observed range.

15.2 Intermediate

15.3 Sales Prediction Analysis

Problem: Sales (Y) vs Ad Spend (X) for $n = 10$. Data: (1,15), (2,18), (3,22), (4,24), (5,28), (6,35), (7,36), (8,42), (9,45), (10,50).
Solution:

(a) **Sum of Squares Calculation:** Given $\sum x = 55, \sum y = 315, \sum x^2 = 385, \sum y^2 = 11163, \sum xy = 2025$. Mean $\bar{x} = 5.5, \bar{y} = 31.5$.

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 385 - \frac{3025}{10} = 385 - 302.5 = 82.5$$
$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 11163 - \frac{99225}{10} = 11163 - 9922.5 = 1240.5$$
$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 2025 - \frac{55 \times 315}{10} = 2025 - 1732.5 = 292.5$$

(b) **Model Estimation:**

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{292.5}{82.5} \approx 3.545$$
$$b_0 = \bar{y} - b_1\bar{x} = 31.5 - (3.545)(5.5) = 31.5 - 19.5 = 12.0$$

Equation: $\hat{y} = 12.0 + 3.55x$.

(c) **Goodness of Fit:**

$$SSR = b_1 S_{xy} = 3.545(292.5) = 1037.05$$
$$R^2 = \frac{SSR}{SST} = \frac{1037.05}{1240.5} \approx 0.836$$
$$r = \sqrt{0.836} \approx 0.914$$

83.6% of the variability in Sales is explained by Ad Spend.

(d) **ANOVA Table:**

$$SSE = S_{yy} - SSR = 1240.5 - 1037.05 = 203.45$$
$$MSR = 1037.05 \quad (df = 1)$$
$$MSE = \frac{SSE}{n-2} = \frac{203.45}{8} \approx 25.43$$
$$F = \frac{MSR}{MSE} = \frac{1037.05}{25.43} \approx 40.78$$

Critical $F_{0.05,1,8} = 5.32$. Since $40.78 > 5.32$, the regression is **Significant**.

(e) **T-test Verification:**

$$SE(b_1) = \sqrt{\frac{MSE}{S_{xx}}} = \sqrt{\frac{25.43}{82.5}} \approx 0.555$$
$$t = \frac{b_1}{SE(b_1)} = \frac{3.545}{0.555} \approx 6.39$$
$$t^2 = (6.39)^2 \approx 40.8 \approx F \quad \text{(Confirmed)}$$

(f) **Prediction at $x = 5.5$:**

$$\hat{y} = 12 + 3.545(5.5) = 31.5$$
$$s = \sqrt{25.43} \approx 5.04$$

Since $x = \bar{x}$, the variance of prediction is minimized ($term(x - \bar{x})^2 = 0$).

$$95\% \text{ CI (Mean)} = \hat{y} \pm t_{\alpha/2} s\sqrt{\frac{1}{n}} = 31.5 \pm 2.306(5.04)\sqrt{0.1} \approx 31.5 \pm 3.67$$
$$95\% \text{ PI (Indiv)} = \hat{y} \pm t_{\alpha/2} s\sqrt{1 + \frac{1}{n}} = 31.5 \pm 2.306(5.04)\sqrt{1.1} \approx 31.5 \pm 12.19$$

$\hat{y} = 12.0 + 3.55x$
$R^2 = 0.836$
$F = 40.78$ (Significant)
Pred $x = 5.5$: 31.5.

15.3 Challenge

15.4 Exponential Growth & Risk

Problem: $Y = \alpha e^{\beta x}$. Data given.
Solution:

(a) **Transformation:** $\ln Y = \ln \alpha + \beta x$. Let $Y' = \ln Y$. x: 1, 2, 3, 4, 5, 6. y': 2.30, 2.71, 3.33, 3.81, 4.41, 5.01.

Using linear regression on (x, y'): $\sum x = 21, \sum y' = 21.57$.

$$S_{xx} = 17.5$$
$$S_{xy'} = 9.565$$
$$b_1'(\beta) = \frac{9.565}{17.5} \approx 0.547$$
$$b_0'(\ln \alpha) = \bar{y}' - b_1'\bar{x} = 3.595 - 0.547(3.5) \approx 1.68$$

Model: $\widehat{\ln Y} = 1.68 + 0.547x$.

(b) **MSE Calculation (Transformed):** Calculate residuals e_i' for log data. $SSE' \approx \sum e_i^2 \approx 0.012$.

$$MSE' = \frac{0.012}{6-2} = 0.003$$
$$s' = \sqrt{0.003} \approx 0.055$$

(c) **Probability (Normal) at $x = 4$:** Predicted mean $\widehat{\ln Y}_4 = 1.68 + 0.547(4) = 3.868$. Threshold: $\ln 50 \approx 3.912$. We need $P(\ln Y > 3.912)$.

$$Z = \frac{3.912 - 3.868}{0.055} = \frac{0.044}{0.055} = 0.8$$
$$P(Z > 0.8) = 1 - 0.7881 = 0.2119$$

There is a 21.2% chance count exceeds 50.

(d) **Probability (Binomial) $n = 10, p = 0.212$:** Let K be number of dishes > 50. $K \sim B(10, 0.212)$.

$$P(K \geq 2) = 1 - [P(K = 0) + P(K = 1)]$$
$$P(K = 0) = (0.788)^{10} \approx 0.092$$
$$P(K = 1) = 10(0.212)(0.788)^9 \approx 0.247$$
$$P(K \geq 2) = 1 - (0.092 + 0.247) = 0.661$$

(a) $\widehat{\ln Y} = 1.68 + 0.55x$
(c) $p \approx 0.212$
(d) $P(K \geq 2) \approx 0.661$

15.5 Proof: BLUE Property

Problem: $b_1 = \sum w_i Y_i$ where $w_i = \frac{x_i - \bar{x}}{S_{xx}}$.
Solution:

(a) **Unbiased:** Properties of w_i: $\sum w_i = 0$ and $\sum w_i x_i = 1$.

$$\begin{aligned} E[b_1] &= E\left[\sum w_i Y_i\right] = \sum w_i E[Y_i] \\ &= \sum w_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum w_i + \beta_1 \sum w_i x_i \\ &= \beta_0(0) + \beta_1(1) = \beta_1 \end{aligned}$$

(b) **Minimum Variance:**

$$\begin{aligned} Var(b_1) &= Var\left(\sum w_i Y_i\right) = \sum w_i^2 Var(Y_i) \quad \text{(Independence)} \\ &= \sigma^2 \sum w_i^2 = \sigma^2 \sum \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 \\ &= \frac{\sigma^2}{S_{xx}^2} \sum (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}^2} \cdot S_{xx} \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

By the Gauss-Markov theorem, this variance is the smallest among all linear unbiased estimators.

$\boxed{\text{Proven.}}$

15.4 Applications

15.6 Interpreting Python Output

Problem: Statsmodels output for Temperature vs Yield.
Solution:

(a) **Model Equation:** From ‘coef‘ column: ‘const‘ = 12.5000, ‘Temperature‘ = 1.8000.

$$\hat{y} = 12.5 + 1.8x$$

(b) **Significance:** Look at ‘P>|t|‘ for Temperature. It is ‘0.000‘. Since $0.000 < 0.01$, Temperature is **Statistically Significant**.

(c) **Confidence Interval [1.184, 2.416]:** This is the 95% Confidence Interval for the **true slope** β_1. We are 95% confident that for every 1 degree increase in temperature, the yield increases by between 1.184 and 2.416 units.

(d) **Normality Assumption:** Jarque-Bera (JB) tests the null hypothesis H_0: Residuals are Normal. ‘Prob(JB): 0.771‘. Since $0.771 > 0.05$, we **Fail to Reject** H_0. The assumption of normality is **valid**.

(a) $\hat{y} = 12.5 + 1.8x$ (b) Significant ($P < 0.01$) (c) CI for Slope (d) Normality Valid ($P_{JB} > 0.05$)

15.7 Interpreting Excel Output

Problem: Standard Excel Regression Table.
Solution:

(a) **Model Fit (R^2):** R Square = 0.8464. The model explains **84.64%** of the variance in Y.

(b) **Error Variance ($\hat{\sigma}^2$):** This is the Mean Square Residual (MS Residual).

$$\hat{\sigma}^2 = 27.04$$

(c) **Standard Error:**

$$SE = \sqrt{MS_{Residual}} = \sqrt{27.04} = 5.2$$

Matches the "Standard Error" row (5.2000).

(d) **F-Test:** Significance F = 0.0000. This means the probability of getting this fit by random chance is zero. The model is **Useful** (Significant).

(a) 84.64% (b) $\hat{\sigma}^2 = 27.04$ (d) Useful (Significant)