Introduction to Linear Regression (Solutions)
STAT-UB.0003: Regression and Forecasting Models

Hypothesis tests (review)

1. We collect a simple random sample of size $n = 100$ from a population. The sample mean is $\bar{x} = 12.4$ and the sample standard deviation is $s = 8.0$. Use this data to test the null hypothesis $H_0 : \mu = 10.0$ against the alternative $H_a : \mu \neq 10.0$, where μ denotes the population mean:

   (a) Compute the test statistic.

   Solution: Since the population standard deviation (σ) is unknown, we use a t-statistic:

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{(12.4) - (10.0)}{(8.0)/\sqrt{(100)}} \\ &= 3.00 \end{aligned}$$

   (b) If the null hypothesis were true and we were to repeat the experiment, we would get a new test statistic. In this hypothetical setting, approximately what is the probability of getting a new test statistic at least as extreme as the observed test statistic we computed in part (a)?

   Solution:
   The t-statistic has $n - 1 = 99$ degrees of freedom. The question asks for

$$\begin{aligned} p &= P(|T_{99}| \geqslant 3.00) \\ &= .0034. \end{aligned}$$

   where T_{99} denotes a random t-statistic with 99 degrees of freedom. To get the value .0034, I used Minitab.
   We can get an approximate probability by using a z table (e.g., Table II from Appendix D). In this case, we get

$$\begin{aligned} p &\approx P(|Z| \geqslant 3.00) \\ &= 1 - 2(.4987) \\ &= .0026. \end{aligned}$$

   (c) What is the p-value for performing this hypothesis test? Give a one-sentence explanation.

   Solution: If the population mean were equal to 10.0, then the chance of seeing data at least as extreme as observed would be $p \approx .0026$.
   A more precise sentence is the following: if the population mean were equal to 10.0 and we were to repeat the experiment—collecting a new sample—then the chance of

seeing a test statistic for the new sample at least as extreme as the test statistic for the observed sample would be $p \approx .0026$.

(d) Using a significance level (α) of 5%, what is the result of the hypothesis test?

Solution: Since $p < .05$, we would reject the null hypothesis at significance level 5%.

Linear regression

2. In the following scenarios, which would you consider to be predictor (x) and which would you consider to be response (y)?
   (a) Sales revenue; Advertising expenditures
   (b) Starting salary after college; Undergraduate GPA
   (c) The current month's sales; the previous month's sales
   (d) The size of an apartment; the sale price of an apartment.
   (e) A restaurant's Zagat Price rating; a restaurant's Zagat Food rating.

> **Solution:** This is a little bit subjective, but the following answers make sense: (a) $y =$ sales revenue; (b) $y =$ starting salary; (c) $y =$ current sales; (d) $y =$ sale price; (e) either makes sense.

3. Let y be the payment (in dollars) for a repair which takes x hours. Suppose that

$$y = 25 + 30\,x.$$

   What is the interpretation of this model?

> **Solution:** There is a positive linear relationship between y and x. Increasing repair time by one hour increases payment by \$30. There is no interpretation for the intercept since repair time is always positive.

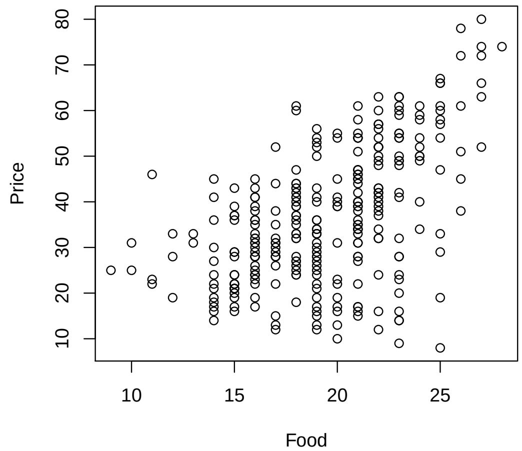
4. Consider two variables measured on 294 restaurants in the 2003 Zagat guide:

$$\begin{aligned} y &= \text{typical dinner price, including one drink and tip (\$)} \\ x &= \text{Zagat quality rating (0–30).} \end{aligned}$$

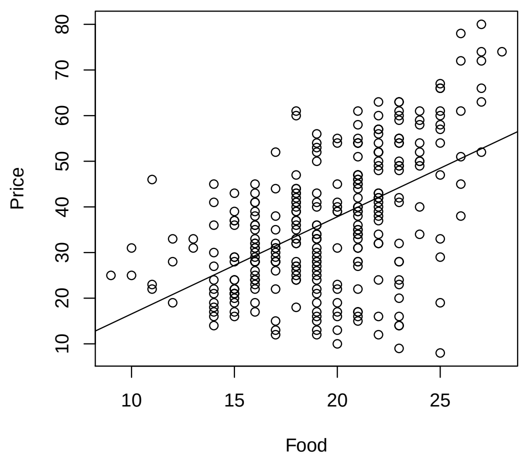
   Here is a scatterplot of y on x:

   Why is an exact linear relationship inappropriate to describe the relationship between y and x?

> **Solution:** There are no values β_0 and β_1 such that $y = \beta_0 + \beta_1 x$ for all restaurants; no straight line fits the data perfectly.



5. Here is the least squares regression fit to the Zagat restaurant data:



Here is the Minitab output from the fit:

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
12.5559  27.93%     27.68%      26.86%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  -4.74     3.95    -1.20    0.232
Food      2.129    0.200    10.64    0.000  1.00

Regression Equation
```

```
Price = -4.74 + 2.129 Food
```

(a) What are the estimated intercept and slope?

> **Solution:** The estimated intercept is $\hat{\beta}_0 = -4.74$; the estimated slope is $\hat{\beta}_1 = 2.129$.

(b) Use the estimated regression model to estimate the average dinner price of all restaurants with a quality rating of 20.

> **Solution:** If Food $= 20$, then estimated expected price per meal ($) is $\widehat{\text{Price}} = -4.74 + 2.129(20) = 37.84$.

(c) In the estimated regression model, what is the interpretation of the slope?

> **Solution:** For every 1-point increase in food quality, the expected dinner price goes up by $2.129.

(d) In the estimated regression model, why doesn't the intercept have a direct interpretation?

> **Solution:** This would be the expected dinner price for a restaurant with a quality of 0. No such restaurant exists (this is outside the range of the data).

6. Refer to the Minitab output from the previous problem, the regression analysis of the Zagat data.

   (a) What is the estimated standard deviation of the error (the "standard error of the regression")? What is the interpretation of this value?

   Solution: The estimated error standard deviation is $s = 12.5559$. Using the empirical rule, the model says that approximately 95% of restaurants have prices within $2s = 25.11$ of the regression line.

   (b) According to the estimated regression model, what is the range of typical prices for restaurants with quality ratings of 20?

   Solution: $37.84 \pm 25.11 = (12.73, 62.95)$

   (c) According to the estimated regression model, what is the range of typical prices for restaurants with quality ratings of 10?

   Solution: In the estimated regression model, when the quality rating is 10, the expected price is $-4.74 + 2.129(10) = 16.55$; the range of typical prices is $16.56 \pm 25.11 = (-8.5441.66)$. Since price can't be negative, we could just as well report the range as $(0, 41.66)$. Note that since $x = 10$ is at the edge of the range of the data, the values predicted by the model are not very reliable.